Sora：大型视觉模型的背景、技术、局限性和机遇综述 【官方论文】

📌 译者：城主

英文论文地址：*https://arxiv.org/pdf/2402.17177v1.pdf*

中文翻译地址：https://mp.weixin.qq.com/s/M9_0oETKIfgJsHTH4t6NdQ

Sora：大型视觉模型的背景、技术、局限性和机遇综述

Yixin Liu1* Kai Zhang1* Yuan Li1* Zhiling Yan1* Chujie Gao1*

Ruoxi Chen1* Zhengqing Yuan1* Yue Huang1*Hanchi Sun1*

Jianfeng Gao2 Lifang He1 Lichao Sun1 †

1利哈伊大学2微软研究院

摘要

Sora是一款由OpenAI在2024年2月发布的文本到视频生成的AI模型。该模型被训练用于根据文本指令生成现实或想象场景的视频，并展现出模拟物理世界的潜力。基于公开的技术报告和逆向工程，本文对该模型的背景、相关技术、应用、现存挑战和文本到视频AI模型的未来方向进行了全面的综述。我们首先追溯Sora的发展，并探究用于构建这一“世界模拟器”的底层技术。然后，我们详细描述了Sora在从电影制作和教育到市场营销等多个行业中的应用和潜在影响。我们讨论了需要解决的主要挑战和局限性，以便广泛部署Sora，例如确保安全且无偏见的视频生成。最后，我们讨论了Sora和一般视频生成模型的未来发展，以及该领域的进步如何能够实现新的人工智能交互方式，提高视频生成的生产力和创造力。

图1：Sora：AI驱动视觉生成的突破。

**平等贡献。顺序是通过掷骰子确定的。Chujie, Ruoxi, Yuan, Yue, 和 Zhengqing 是在利哈伊大学LAIR实验室的访问学生。GitHub链接为 https://github.com/lichao-sun/SoraReview*

† Lichao Sun 是共同通讯作者：lis221@lehigh.edu

目录

1 引言

自从 2022 年 11 月 ChatGPT 发布以来，人工智能技术的出现标志着一个重大的转变，重塑了交互方式，并深入融入日常生活和行业的各个方面[1, 2]。基于这一势头，OpenAI 在 2024 年 2 月发布了

Sora，一个文本到视频的生成式 AI 模型，能够根据文本提示生成现实或想象场景的视频。与之前的视频生成模型相比，Sora 的特点是能够在遵循用户文本指令的同时，生成长达 1 分钟的高质量视频[3]。Sora 的进步体现了长期以来人工智能研究任务的实质，即赋予 AI 系统（或 AI 代理）理解和与运动中的物理世界互动的能力。这涉及到开发不仅能解释复杂用户指令，而且能将这种理解应用于通过动态和富有上下文的模拟解决现实世界问题的 AI 模型。

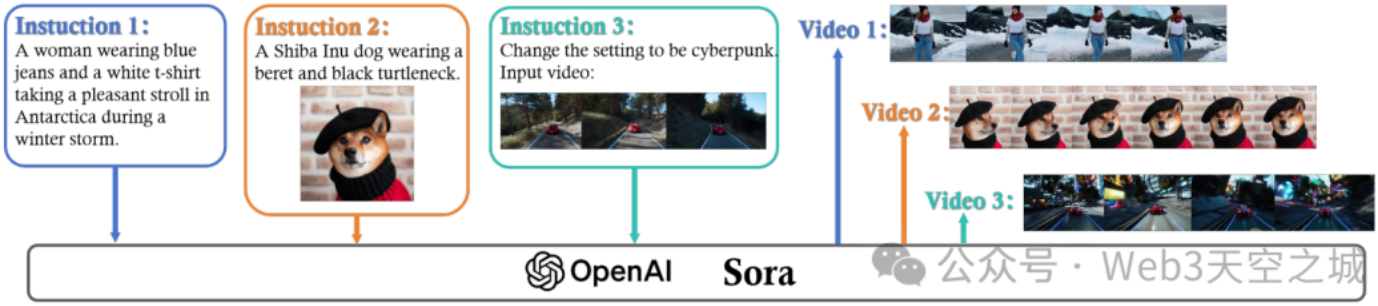


图 2：Sora 在文本到视频生成中的示例。文本指令被给予 OpenAI Sora 模型，它根据指令生成三个视频。

Sora 展示了准确解释和执行复杂人类指令的显著能力，如图 2 所示。该模型能生成包含多个执行特定动作的角色以及复杂背景的详细场景。研究人员将 Sora 的熟练程度归因于不仅处理用户生成的文本提示，而且还能辨别场景内元素之间复杂的相互作用。Sora 最引人注目的方面之一是其生成长达一分钟的视频的能力，同时保持高视觉质量和引人入胜的视觉连贯性。与只能生成短视频片段的早期模型不同，Sora 的一分钟长视频创作具有进展感和从第一帧到最后一帧的视觉一致性之旅。此外，Sora 的进步在于其生成具有细腻运动和互动描绘的扩展视频序列的能力，克服了早期视频生成模型所特有的短片段和简单视觉呈现的限制。这一能力代表了 AI 驱动创意工具向前的一大步，允许用户将文本叙述转换为丰富的视觉故事。总的来说，这些进步展示了 Sora 作为世界模拟器的潜力，为描绘场景的物理和上下文动态提供了细腻的见解。[3]。

技术。Sora 的核心是一个预训练的扩散变换器[4]。变换器模型已被证明对许多自然语言任务具有可扩展性和有效性。与 GPT-4 等强大的大型语言模型（LLMs）类似，Sora 能够解析文本并理解复杂的用户指令。为了使视频生成在计算上高效，Sora 使用时空潜在补丁作为其构建块。具体来说，Sora 将原始输入视频压缩成一个时空潜在表示。然后，从压缩视频中提取一系列时空潜在补丁，以封装短时间间隔内的视觉外观和运动动态。这些补丁，类似于语言模型中的单词标记，为 Sora 提供了用于构建视频的详细视觉短语。Sora 的文本到视频生成是通过扩散变换器模型执行的。从一个充满视觉噪声的帧开始，模型迭代地去除噪声并根据提供的文本提示引入特定细节。本质上，生成的视频通过多步精炼过程出现，每一步都使视频更加符合期望的内容和质量。

Sora 的亮点。Sora 的能力在各个方面都有深远的影响：

· **提高模拟能力：**将 Sora 进行大规模训练归功于其模拟物理世界各个方面的显著能力。尽管缺乏显式的 3D 建模，Sora 展现了具有动态相机运动和长距离连贯性的 3D 一致性，包括对象持久性和模拟与世界的简单互动。此外，Sora 有趣地模拟了像 Minecraft 这样的数字环境，由基本策略控制，同时保持视觉保真度。这些涌现能力表明，扩大视频模型的规模在创建模拟物理和数字世界复杂性的 AI 模型方面是有效的。

· **促进创造力：**想象通过文本概述一个概念，无论是一个简单的对象还是一个完整的场景，并在几秒钟内看到一个现实的或高度风格化的视频被渲染出来。Sora 使设计过程加速，加快了探索和精炼想法的速度，从而显著提高了艺术家、电影制作人和设计师的创造力。

· **推动教育创新：**视觉辅助工具长期以来一直是理解教育中重要概念的核心。有了 Sora，教育工作者可以轻松地将课堂计划从文本转换为视频，以吸引学生的注意力并提高学习效率。从科学模拟到历史剧情再现，可能性是无限的。

· **增强可访问性：**在视觉领域增强可访问性至关重要。Sora 通过将文本描述转换为视觉内容提供了一种创新解决方案。这一能力使所有人，包括视觉障碍者，都能积极参与内容创作并以更有效的方式与他人互动。因此，它为每个人提供了通过视频表达自己想法的机会，创造了一个更具包容性的环境。

· **培育新兴应用：**Sora 的应用范围广泛。例如，营销人员可能使用它创建针对特定受众描述的动态广告。游戏开发者可能使用它从玩家叙述中生成定制化的视觉效果甚至角色动作。

限制和机遇。尽管 Sora 的成就突出了 AI 方面的重大进步，但仍存在挑战。描绘复杂动作或捕捉微妙的面部表情是模型可以增强的领域之一。此外，如何减轻生成内容中的偏见并防止有害视觉输出等伦理考虑，强调了开发者、研究人员和更广泛社区负责任使用的重要性。确保 Sora 的输出始终安全且无偏见是一个主要挑战。视频生成领域正在迅速发展，学术和行业研究团队正在不懈努力。竞争性文本到视频模型的出现表明，Sora 可能很快就会成为一个动态生态系统的一部分。这种协作和竞争环境促进了创新，导致视频质量的提高和新应用的出现，这些应用有助于提高工人的生产力并使人们的生活更加有趣。

我们的贡献。基于已发布的技术报告和我们的逆向工程，本文提出了 Sora 的背景、相关技术、新兴应用、当前限制和未来机遇的首次全面审查。

2 背景

2.1 历史

在计算机视觉（CV）领域，在深度学习革命之前，传统的图像生成技术依赖于诸如纹理合成[5]和纹理映射[6]等方法，这些方法基于手工制作的特征。然而，这些方法在生成复杂和生动的图像方面的能力是有限的。

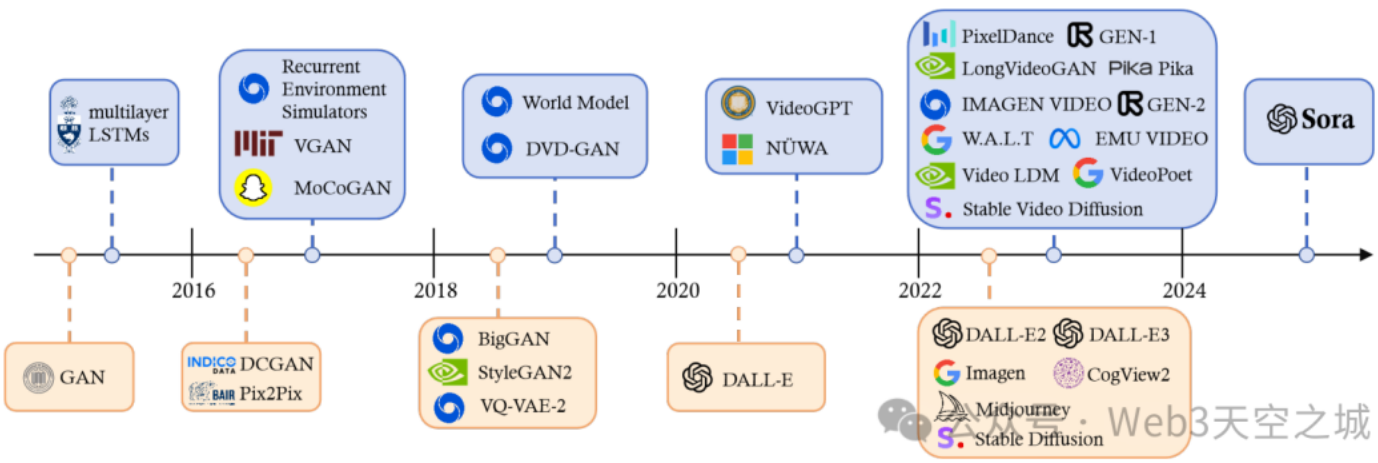


图3：视觉领域生成式AI的历史。

生成对抗网络（GANs）[7]和变分自编码器（VAEs）[8]的引入标志着一个重要的转折点，因为它在各种应用中展现出了非凡的能力。随后的发展，如流模型[9]和扩散模型[10]，进一步提高了图像生成的细节和质量。最近在人工智能生成内容（AIGC）技术方面的进展，使内容创作民主化，使用户能够通过简单的文本指令生成所需内容[11]。

在过去的十年中，生成式计算机视觉（CV）模型的发展采取了多种路径，如图3所示。这一格局在成功应用变压器架构[12]于自然语言处理（NLP）后开始显著转变，如BERT[13]和GPT[14]所示。在CV中，研究人员通过将变压器架构与视觉组件相结合，将这一概念推向更远，使其能够应用于下游CV任务，如视觉变压器（ViT）[15]和Swin变压器[16]。与变压器的成功平行，扩散模型在图像和视频生成领域也取得了重大进展[10]。扩散模型提供了一个数学上合理的框架，通过U-Nets[17]将噪声转换成图像，其中U-Nets通过学习在每一步预测和减轻噪声来促进这一过程。

自2021年以来，AI领域的一个重要焦点是能够解释人类指令的生成式语言和视觉模型，即多模态模型。例如，CLIP[18]是一种开创性的视觉-语言模型，它将变压器架构与视觉元素相结合，使其能够在大量的文本和图像数据集上进行训练。通过从一开始就整合视觉和语言知识，CLIP可以在多模态生成框架中充当图像编码器。另一个值得注意的例子是Stable Diffusion[19]，这是一个多才多艺的文本到图像AI模型，以其适应性和易用性而受到赞誉。它采用变压器架构和潜在扩散技术来解码文本输入，并生成各种风格的图像，进一步展示了多模态AI的进步。

随着2022年11月ChatGPT的发布，我们在2023年见证了商业文本到图像产品的出现，如Stable Diffusion[19]、Midjourney[20]、DALL-E 3[21]。这些工具使用户能够用简单的文本提示生成高分辨率和高质量的新图像，展示了AI在创意图像生成方面的潜力。然而，从文本到图像过渡到文本到视频由于视频的时间复杂性而具有挑战性。尽管工业界和学术界做出了许多努力，但大多数现有的视频生成工具，如Pika[22]和Gen-2[23]，仅限于生成几秒钟的短视频片段。在这种背景下，Sora代表了一个重大突破，类似于ChatGPT在NLP领域的影响。Sora是第一个能够根据人类指令生成长达一分钟的视频的模型，标志着对生成式AI研究和开发产生深远影响的里程碑。为了便于轻松访问最新的视觉生成模型进展，最新的作品已被汇编并提供在附录和我们的GitHub中。

2.2 高级概念

视觉模型的规模化定律。有了LLMs的规模化定律，自然会问视觉模型的发展是否遵循类似的规模化定律。最近，Zhai等人[24]展示了，有足够训练数据的ViT模型的性能-计算前沿大致遵循（饱和）幂律。继他们之后，谷歌研究[25]提出了一种高效稳定训练22B参数ViT的方法。结果显示，使用冻结模型产生嵌入，然后在顶部训练薄层可以实现出色的性能。Sora作为一个大型视觉模型（LVM），符合这些规模化原则，揭示了文本到视频生成中的几种新兴能力。这一重大进展强调了LVMs实现类似LLMs所见进步的潜力。

新兴能力。LLMs中的新兴能力是在某些规模上——通常与模型参数的大小有关——表现出的复杂行为或功能，这些行为或功能并未被开发者明确编程或预期。这些能力被称为“新兴”，因为它们源于模型在多样化数据集上的全面训练，以及其庞大的参数数量。这种组合使模型能够形成联系并做出超越简单模式识别或死记硬背的推断。通常，这些能力的出现不能通过从小规模模型的性能外推来直接预测。虽然许多LLMs，如ChatGPT和GPT-4，展示了新兴能力，但直到Sora的出现，展示类似能力的视

觉模型还很少。根据Sora的技术报告，它是第一个展示确认新兴能力的视觉模型，标志着计算机视觉领域的一个重要里程碑。

除了其新兴能力，Sora还展示了其他显著能力，包括遵循指令、视觉提示工程和视频理解。Sora的这些功能方面代表了视觉领域的重大进步，并将在后续部分进行探讨和讨论。

3 技术

3.1 Sora概述

从核心本质上看，Sora是一个具有灵活采样维度的扩散变压器[4]，如图4所示。它有三个部分：（1）时空压缩器首先将原始视频映射到潜在空间。（2）ViT然后处理标记化的潜在表示，并输出去噪的潜在表示。（3）类似CLIP[26]的条件机制接收LLM增强的用户指令和可能的视觉提示，以指导扩散模型生成风格化或主题化的视频。经过多次去噪

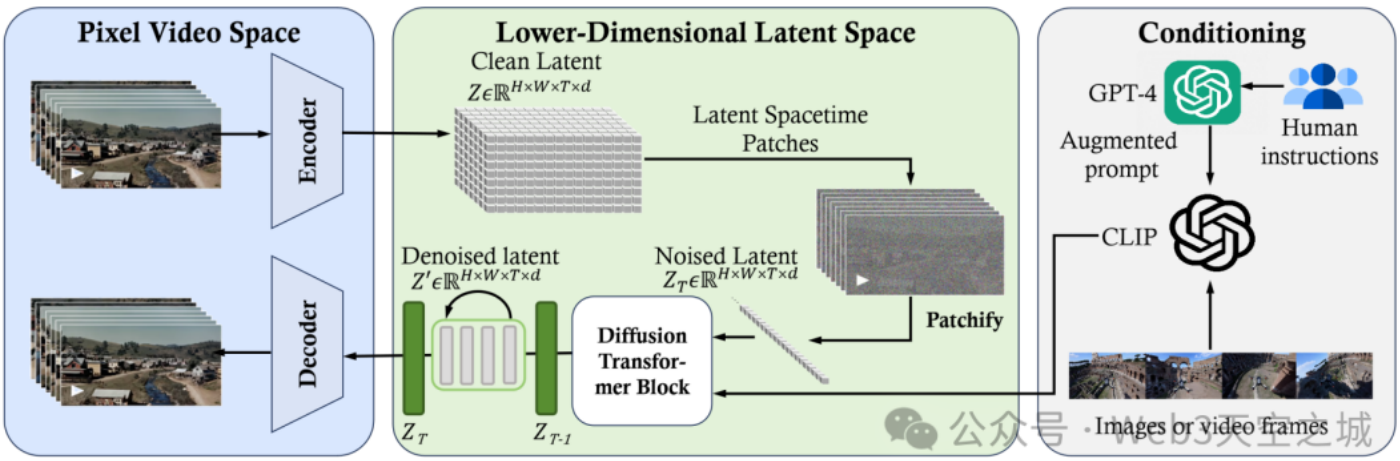


图4：逆向工程：Sora框架概览

在这一步骤中，生成视频的潜在表示被获得，然后通过相应的解码器映射回像素空间。在本节中，我们的目标是对Sora使用的技术进行逆向工程，并讨论广泛的相关工作。

3.2 数据预处理

3.2.1 变化的持续时间、分辨率、宽高比

Sora的一个区别特征是其能够在原始尺寸上训练、理解和生成视频及图像，如图5所示。传统方法通常会调整视频的大小、裁剪或调整宽高比，以适应一个统一标准——通常是以固定低分辨率的正方形帧的短片段[27][28][29]。这些样本通常在更宽的时间跨度上生成，并依赖于分别训练的帧插入和分辨率渲染模型作为最后一步，这在视频中造成了不一致性。利用扩散变换器架构[4]（见3.2.4节），Sora是第一个拥抱视觉数据多样性的模型，并且可以在从宽屏1920x1080p视频到竖屏1080x1920p视频以及之间的任何格式上采样，而不会损害它们的原始尺寸。

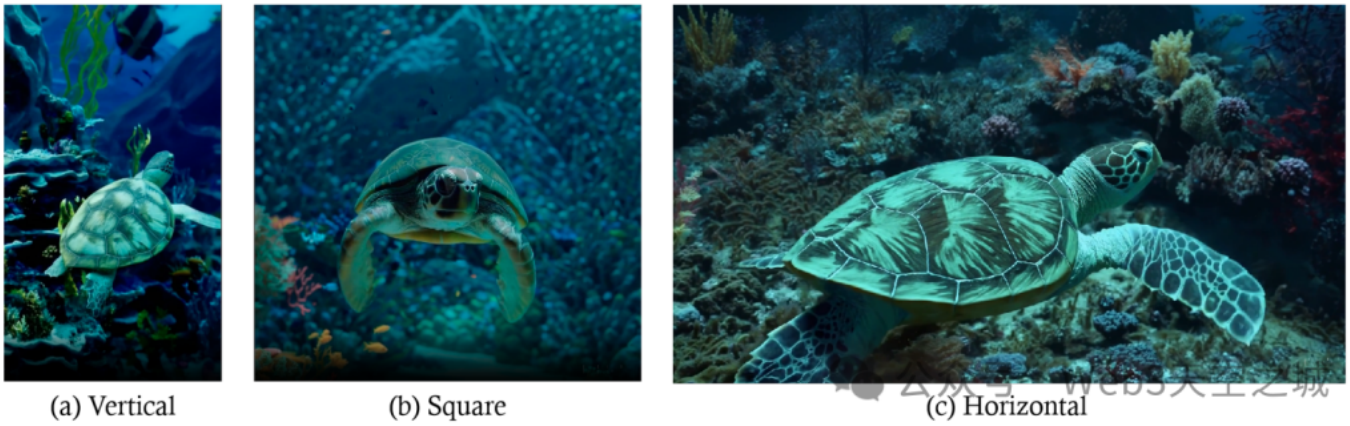


图5：Sora可以生成从1920x1080p到1080x1920p及其间任何尺寸或分辨率的图像。

图6：Sora（右）与一个修改版的模型（左）之间的比较，后者将视频裁剪成正方形——这是模型训练中的一种常见做法——凸显了优势。

在原始尺寸上训练数据显著改善了生成视频的构图和框架。经验发现，通过保持原始宽高比，Sora实现了更自然和连贯的视觉叙事。如图6所示，Sora与一个在统一裁剪的正方形视频上训练的模型之间的比较展示了明显的优势。Sora生成的视频展示了更好的框架，确保场景中完全捕捉到了主体，与正方形裁剪导致的有时被截断的视图相反。

这种对原始视频和图像特征的细腻理解和保留标志着生成模型领域的一个重大进步。Sora的方法不仅展示了生成更真实和吸引人视频的潜力，而且还强调了在训练数据中的多样性对于在生成AI中实现高质量结果的重要性。Sora的训练方法与Richard Sutton的《苦涩的教训》[30]的核心原则一致，该原则指出，利用计算而不是人为设计的特征，可以导致更有效和灵活的AI系统。正如扩散变换器的原始设计寻求简单性和可扩展性[31]，Sora在原始尺寸上训练数据的策略避免了传统AI对人为抽象的依赖，而是选择了一种随着计算能力扩展的通用方法。在本节的其余部分，我们尝试对Sora的架构设计进行逆向工程，并讨论实现这一惊人特性的相关技术。

3.2.2 统一的视觉表示

为了有效处理包括具有不同持续时间、分辨率和宽高比的图像和视频在内的多样化视觉输入，一个关键方法涉及将所有形式的视觉数据转换为统一表示，这有助于生成模型的大规模训练。具体来说，Sora通过最初将视频压缩到较低维度的潜在空间，然后将表示分解为时空补丁来对视频进行分块。然而，Sora的技术报告[3]仅提出了一个高层次的想法，使得研究社区难以复制。在本节中，我们尝试逆向工程潜在的成分和技术路径。此外，我们将讨论可行的替代方案，这些方案可以复制Sora的功能，借鉴现有文献中的见解。

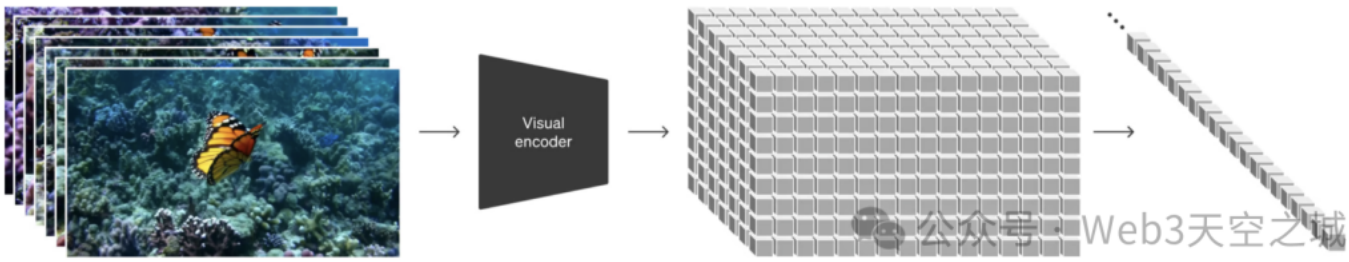


图7：在高层次上，Sora通过首先将视频压缩到较低维度的潜在空间，然后将表示分解为时空补丁来对视频进行分块。来源：Sora的技术报告[3]。

3.2.3 视频压缩网络

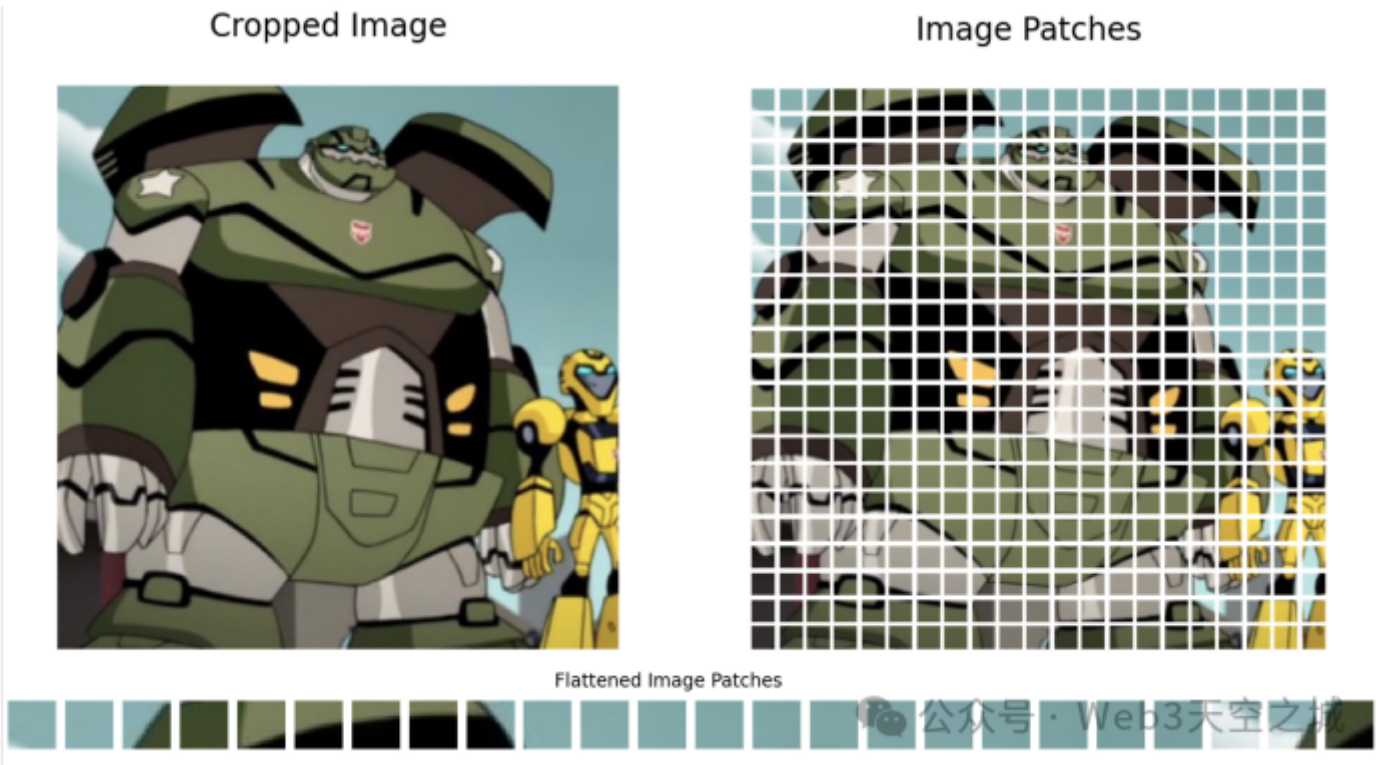


图8：ViT将图像分割成固定大小的补丁，线性嵌入每个补丁，添加位置嵌入，然后将生成的向量序列输入标准的Transformer编码器。

Sora的视频压缩网络（或视觉编码器）旨在降低输入数据的维度，尤其是原始视频，并输出一个在时间和空间上都被压缩的潜在表示，如图7所示。根据技术报告中的参考文献，压缩网络是基于VAE或向量量化-VAE（VQ-VAE）[32]构建的。然而，如果不使用调整大小和裁剪，如技术报告中所述，VAE难

以将任何大小的视觉数据映射到统一且固定大小的潜在空间。我们总结了两种不同的实现来解决这个问题：**空间补丁压缩。**

这涉及将视频帧转换为固定大小的补丁，类似于ViT[15]和MAE[33]（见图8）中使用的方法，然后将它们编码到潜在空间中。这种方法特别适用于适应不同分辨率和宽高比的视频，因为它通过处理个别补丁来编码整个帧。随后，这些空间标记按时间顺序组织，创建一个空间-时间潜在表示。这种技术强调了几个关键考虑因素：时间维度的可变性——鉴于训练视频的持续时间不同，时间维度潜在空间表示的维度不能固定。为了解决这个问题，可以选择采样特定数量的帧（对于较短的视频可能需要填充或时间插值[34]），或者为后续处理定义一个通用的扩展（超长）输入长度（更多细节在第3.2.4节中描述）；使用预训练的视觉编码器——为了处理高分辨率视频，利用现有的预训练视觉编码器，如Stable Diffusion[19]中的VAE编码器，对大多数研究者来说是可取的，而Sora的团队则预期将从头开始通过训练潜在扩散模型[19, 35, 36]的方式训练自己的压缩网络及解码器（视频生成器）。这些编码器可以高效压缩大尺寸的图块（例如，256×256），便于大规模数据的管理；时间信息聚合——由于这种方法主要关注空间图块压缩，因此需要一个额外的机制在模型内聚合时间信息。这一方面对于捕捉随时间动态变化至关重要，并在后续章节中进一步阐述（详见第3.2.6节和图14）。

空间-时间-图块压缩。这项技术旨在封装视频数据的空间和时间维度，提供全面的表示。这项技术不仅仅分析静态帧，还考虑帧间的运动和变化，从而捕捉视频的动态方面。使用3D卷积作为实现这种整合的直接和强大方法[37]。图形说明和与纯空间图块化的比较在图9中展示。与空间图块压缩类似，采用空间-时间图块压缩并预设卷积核参数——如固定的核大小、步长和输出通道——会由于视频输入的不同特性导致潜在空间维度的变化。这种变异主要由处理的视频的不同持续时间和分辨率驱动。为了缓解这一挑战，采用空间图块化的方法同样适用且有效。

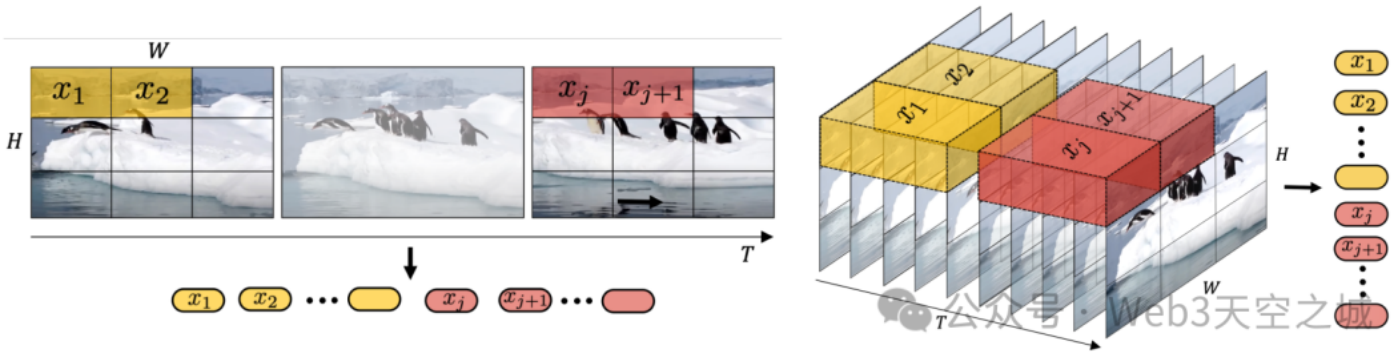


图9：视频压缩的不同图块化比较。来源：ViViT[38]。（左）空间图块化仅采样nt帧，并独立嵌入每个2D帧，遵循ViT。（右）空间-时间图块化提取并线性嵌入跨越时空输入体积的非重叠或重叠管状体。

总结来说，我们基于VAE或其变体如VQ-VQE对两种图块级压缩方法进行了逆向工程，因为对图块的操作在处理不同类型的视频时更加灵活。由于Sora旨在生成高保真视频，因此使用大图块尺寸或核大小进行高效压缩。这里，我们期望使用固定大小的图块，以简化、可扩展性和训练稳定性。但也可以使用不同大小的图块[39]，使整个帧或视频在潜在空间的维度一致。然而，这可能导致无效的位置编码，并为解码器生成不同大小潜在图块的视频带来挑战。

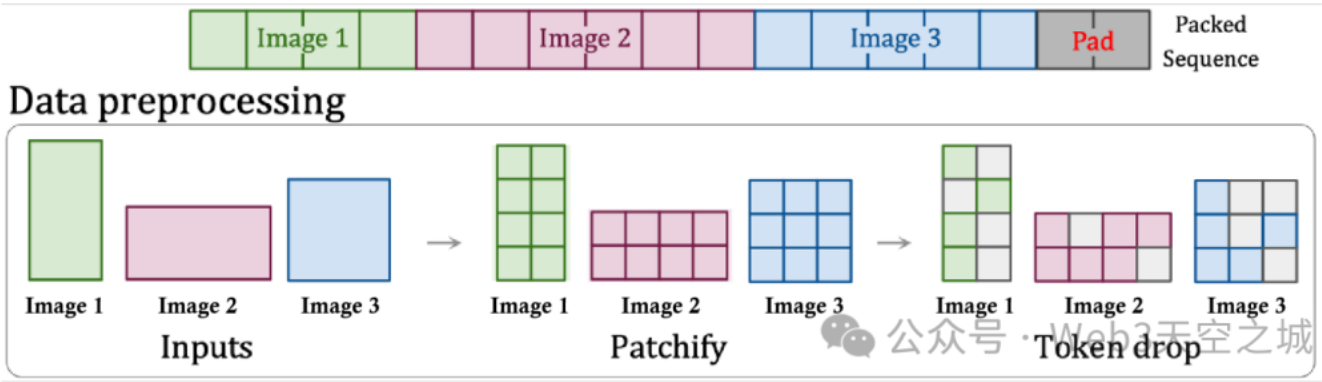


图10：图块打包允许保持纵横比的可变分辨率图像或视频。6令牌丢弃在某种程度上可以被视为数据增强。来源：NaViT[40]。

3.2.4 时空潜在图块

在压缩网络部分仍有一个关键问题：如何处理潜在空间维度的变化性（即，来自不同视频类型的潜在特征块或图块的数量）在将图块输入到扩散变换器的输入层之前。这里，我们讨论了几种解决方案。基于Sora的技术报告和相应的参考文献，图块打包（PNP）[40]可能是解决方案。PNP将来自不同图像的多个图块打包成单个序列，如图10所示。这种方法受到自然语言处理中示例打包的启发[41]，通过丢弃令牌来适应可变长度输入的高效训练。这里需要在压缩网络中完成图块化和令牌嵌入步骤，但Sora可能会进一步将潜在图块化为变换器令牌，如扩散变换器所做[4]。无论是否进行第二轮图块化，我们需要解决两个问题，如何以紧凑的方式打包这些令牌以及如何控制应该丢弃哪些令牌。对于第一个问题，使用了一种简单的贪婪方法，将示例添加到有足够剩余空间的第一个序列中。一旦没有更多示例可以适配，序列就用填充令牌填充，产生批处理操作所需的固定序列长度。这种简单的打包算法可能导致显著的填充，这取决于输入长度的分布。另一方面，我们可以通过调整序列长度和限制填充来控制我们采样的分辨率和帧，以确保高效打包。对于第二个问题，一种直观的方法是丢弃相似的令牌[42, 43, 33, 44]，或者像PNP那样应用丢弃率调度器。然而，值得注意的是，3D一致性是Sora的一个好特性。在训练期间丢弃令牌可能会忽略细微的细节。因此，我们相信OpenAI可能会使用超长的上下文窗口并打包所有视频的令牌，尽管这样做在计算上是昂贵的，例如，多头注意力[45, 46]操作在序列长度上展示出二次成本。具体来说，来自长时视频的时空潜在图块可以打包成一个序列，而来自几个短时视频的图块则在另一个序列中连接。

3.2.5 讨论

我们讨论了Sora可能使用的两种数据预处理的技术解决方案。这两种解决方案都是在图块级别执行的，因为它们对于建模具有灵活性和可扩展性的特点。与之前将视频调整大小、裁剪或修剪到标准大小的方法不同，Sora在其原生大小上进行训练。尽管有几个好处（详见第3.2.1节分析），但它带来了一些技术挑战，其中最显著的是，神经网络不能固有地处理具有可变持续时间、分辨率和纵横比的视觉数据。通过逆向工程，我们相信Sora首先将视觉图块压缩成低维潜在表示，并将这样的潜在图块或进一步图块化的潜在图块排列成序列，然后将噪声注入这些潜在图块中。在将它们输入扩散变换器的输入层之前。Sora采用了时空分块化，因为它易于实现，并且可以通过高信息密度的令牌有效减少上下文长度，并降低后续对时间信息建模的复杂性。对于研究社区，我们推荐使用成本效益高的替代解决方案进行视频压缩和表示，包括利用预训练的检查点（例如，压缩网络）[47]，缩短上下文窗口，使

用轻量级建模机制如（分组的）多查询注意力[48, 49]或高效的架构（例如，Mamba [50]），必要时对数据进行下采样和丢弃令牌。视频建模的效果与效率之间的权衡是一个重要的研究课题。

3.2.6 扩散变换器

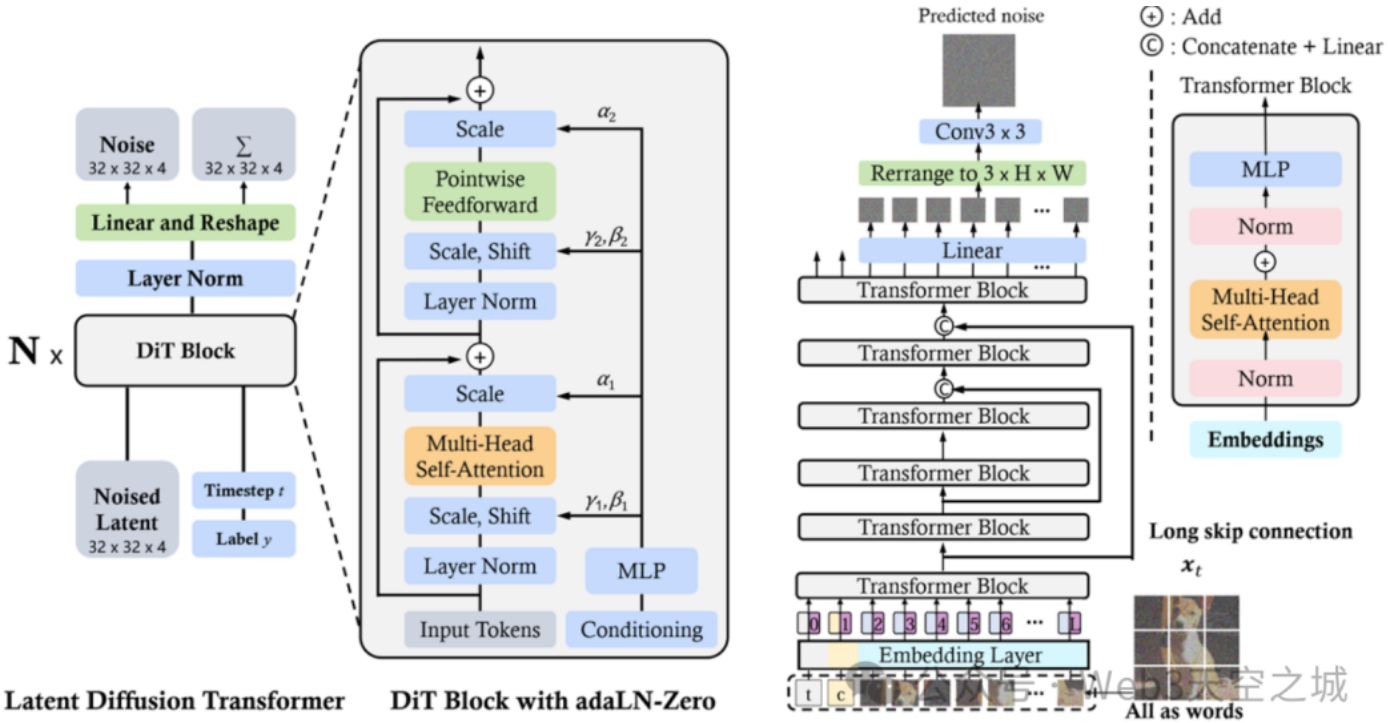


图11：DiT（左）和U-ViT（右）的整体框架。

3.3 建模

图像扩散变换器。传统的扩散模型[51, 52, 53]主要利用包括下采样和上采样块的卷积U-Net作为去噪网络骨干。然而，最近的研究表明，U-Net架构对扩散模型的良好性能并不是关键。通过整合更灵活的变换器架构，基于变换器的扩散模型可以使用更多的训练数据和更大的模型参数。沿着这一思路，DiT[4]和U-ViT[54]是首批采用视觉变换器的潜在扩散模型之一。如同ViT，DiT采用了多头自注意力层和逐点前馈网络，其中穿插了一些层归一化和缩放层。此外，如图11所示，DiT通过自适应层归一化（AdaLN）加上一个额外的MLP层进行条件化，该层用于零初始化，它将每个残差块初始化为恒等函数，从而极大地稳定了训练过程。DiT的可扩展性和灵活性得到了实证验证。DiT成为扩散模型的新骨干。在U-ViT中，如图11所示，它们将所有输入，包括时间、条件和噪声图像块，视为令牌，并在浅层和深层变换器层之间提出长跳跃连接。结果表明，基于CNN的U-Net中的下采样和上采样操作并不总是必要的，U-ViT在图像和文本到图像生成中取得了破纪录的FID分数。

像掩码自编码器（MAE）[33]一样，掩码扩散变换器（MDT）[55]将掩码潜在建模整合到扩散过程中，以显式增强图像合成中对象语义部分之间的上下文关系学习。具体来说，如图12所示，MDT在训练期间使用侧插值来进行额外的掩码令牌重建任务，以提高训练效率并学习强大的上下文感知位置嵌入以用于推理。与DiT[4]相比，MDT实现了更好的性能和更快的学习速度。Hatamizadeh等人[56]介绍了扩散视觉变换器（DiffiT），它使用了一个时间依赖的自注意力（TMSA）模块来对采样时间步骤上的动态去噪行为进行建模，而不是使用AdaLN（即，移位和缩放）进行时间条件建模。此外，DiffiT使用了

两种混合的分层架构，分别用于像素空间和潜在空间中的高效去噪，并在各种生成任务中实现了新的最佳结果。总的来说，这些研究在使用视觉变换器进行图像潜在扩散方面展示了有希望的结果，为未来其他模态的研究铺平了道路。

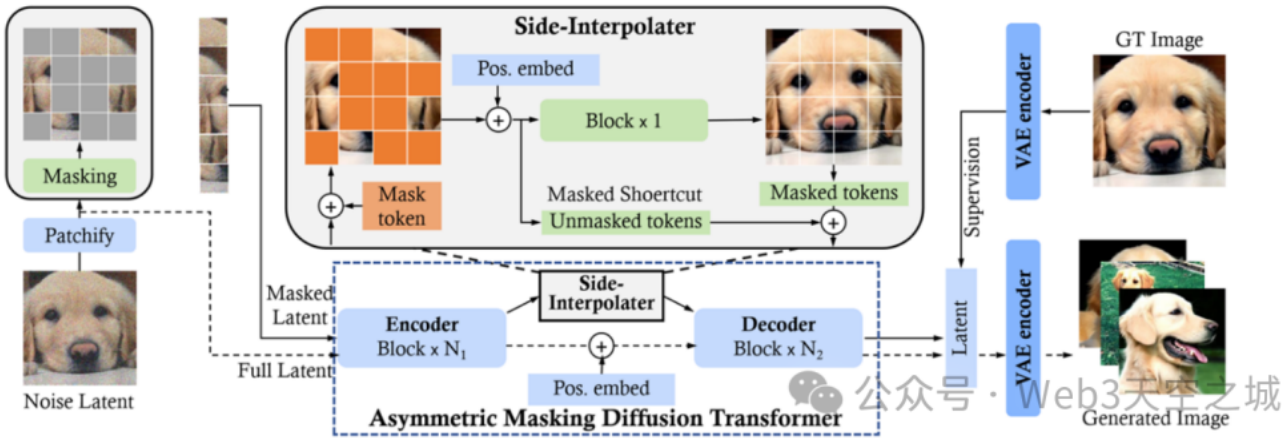
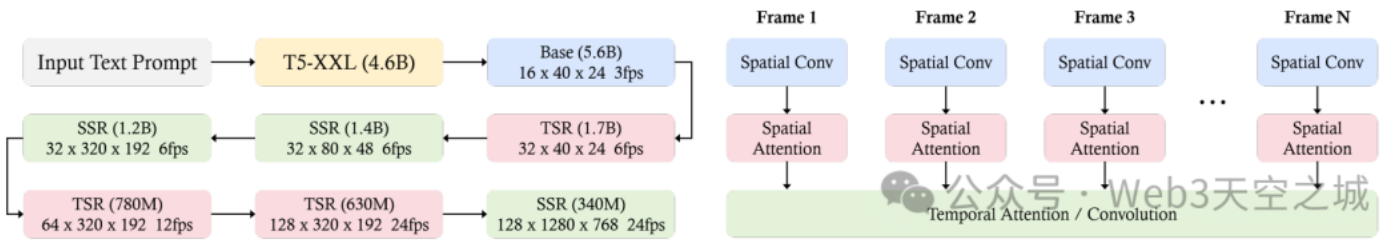


图12：掩码扩散变换器（MDT）的整体框架。实线/虚线表示每个时间步骤的训练/推理过程。掩码和侧插值仅在训练期间使用，在推理期间被移除。

视频扩散变换器。在文本到图像（T2I）扩散模型的基础工作之上，最近的研究集中在实现扩散变换器用于文本到视频（T2V）生成任务的潜力。由于视频的时间性质，将DiTs应用于视频领域的关键挑战是：i) 如何在潜在空间中对视频进行空间和时间上的压缩以实现高效去噪；ii) 如何将压缩的潜在表示转换为块并将它们输入变换器；以及iii) 如何处理长期的时间和空间依赖性并确保内容一致性。请参阅第3.2.3节以了解第一个挑战。在本节中，我们将讨论旨在在空间和时间压缩的潜在空间中操作的基于变换器的去噪网络架构，我们将详细回顾OpenAI Sora技术报告参考列表中描述的两项重要工作（Imagen Video [29]和Video LDM [36]）。



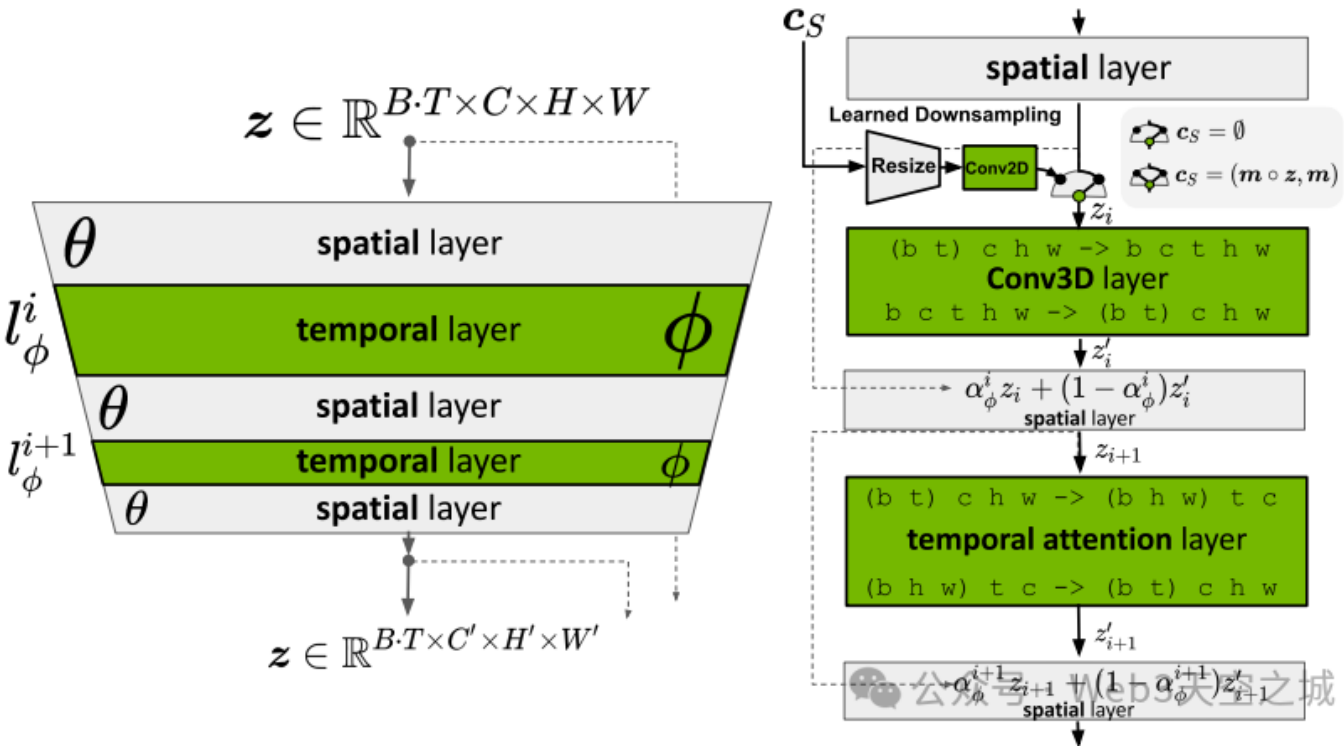
(a)左：级联扩散模型。*级联采样管道的基础扩散模型和六个向上采样模型，操作的空间和时间。文本嵌入被注入到所有的扩散模型。*

(b)右：视频U-Net时空可分离块。*空间操作进行独立的帧与共享的参数，而时间的操作混合激活帧。时间注意仅用于在基本模型的记忆效率。*

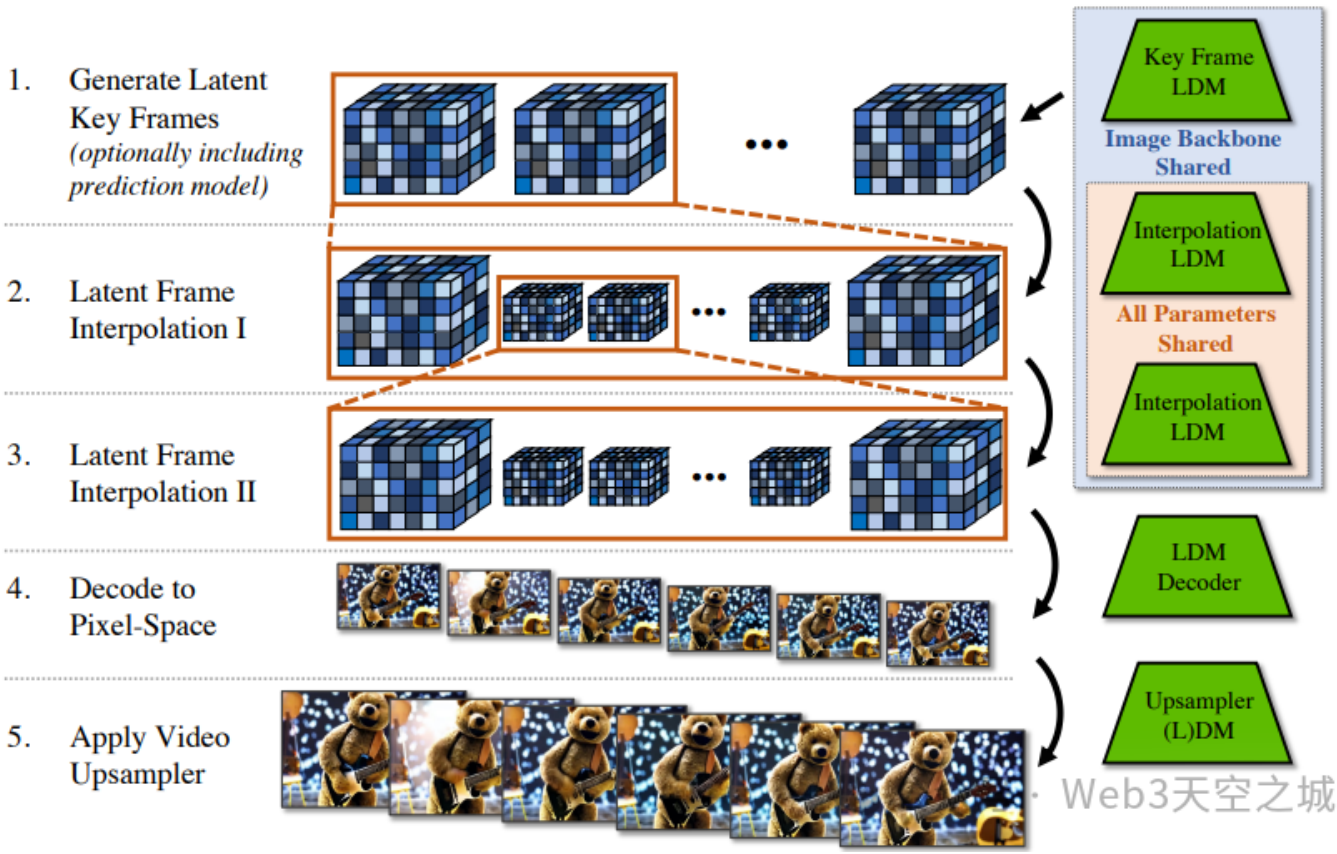
图13：Imagen Video的整体框架。来源：Imagen Video [29]。

基础模型用于低分辨率视频生成，然后通过级联扩散模型进行细化以提高分辨率。基础视频和超分辨率模型采用3D U-Net架构，以时空可分离的方式使用。该架构将时间注意力和卷积层与空间对应物结合起来，以有效捕获帧间依赖性。它采用v-预测参数化以确保数值稳定性，并使用条件增强来促进模型间的并行训练。该过程涉及对图像和视频的联合训练，将每个图像视为一个帧以利用更大的数据集，并使用无分类器引导[57]来增强提示的忠实度。应用渐进式蒸馏[58]来简化采样过程，显著减少计算负

载同时保持感知质量。结合这些方法和技术，Imagen Video不仅能生成高保真度的视频，还能展现出卓越的可控性，如其生成多样化视频、文字动画和各种艺术风格内容的能力所示。



14(a)额外的时间层。通过插入学习将帧对齐成时间一致序列的时间层，将预训练的LDM转变为视频生成器。在优化过程中，图像骨干θ保持固定，只有时间层li的参数 ϑ 接受训练。



14(b)视频LDM堆栈。视频LDM首先生成稀疏关键帧，然后使用相同的潜在扩散模型进行两次时间插值以实现高帧率。最后，将潜在视频解码到像素空间，并可选地应用视频上采样器扩散模型。

图14：视频LDM的整体框架。来源：视频LDM [36]。

Blattmann等人[36]提出将2D潜在扩散模型转变为视频潜在扩散模型（视频LDM）。他们通过在U-Net骨干和VAE解码器的现有空间层之间添加一些事后时间层来实现这一点，这些时间层学习对齐单个帧。这些时间层在编码的视频数据上训练，而空间层保持固定，使模型能够利用大型图像数据集进行预训练。LDM的解码器经过微调以在像素空间中实现时间一致性，并且时间对齐扩散模型上采样器用于增强空间分辨率。为了生成非常长的视频，模型被训练以预测给定一些上下文帧的未来帧，允许在采样过程中使用无分类器引导。为了实现高时间分辨率，视频合成过程被划分为关键帧生成和这些关键帧之间的插值。遵循级联LDMs，使用DM进一步将视频LDM输出扩大4倍，确保高空间分辨率同时保持时间一致性。这种方法使得以计算高效的方式生成全局一致的长视频成为可能。此外，作者展示了将预训练的图像LDM（例如，稳定扩散）转换为文本到视频模型的能力，仅通过训练时间对齐层，实现了高达1280×2048分辨率的视频合成。

3.3.1 讨论

空间和时间上采样的级联扩散模型。Sora能生成高分辨率视频。通过回顾现有工作和我们的逆向工程，我们推测Sora也利用了级联扩散模型架构[59]，该架构由一个基础模型和多个时空细化模型组成。在基础扩散模型和低分辨率扩散模型中，考虑到在高分辨率情况下使用注意力机制的高计算成本和有限的性能增益，注意力模块不太可能被大量使用。对于空间和时间场景一致性，正如先前的工作所示，对于视频/场景生成，时间一致性比空间一致性更重要，Sora可能会通过使用较长视频（用于时间一致性）和较低分辨率的高效训练策略来实现这一点。此外，Sora可能会使用v参数化扩散模型[58]，考虑到其与预测原始潜在值或噪声⊠的其他变体相比的卓越性能。

关于潜在编码器。为了训练效率，大多数现有工作利用稳定扩散[60, 61]的预训练VAE编码器作为初始化模型检查点。然而，编码器缺乏时间压缩能力。尽管一些工作提出只对解码器进行微调以处理时间信息，但解码器在压缩潜在空间中处理视频时间数据的性能仍然不是最佳的。根据技术报告，我们的逆向工程显示，与其使用现有的预训练VAE编码器，Sora可能使用从头开始在视频数据上训练的时空VAE编码器，该编码器在具有视频导向的压缩潜在空间中的性能优于现有的编码器。

3.4 语言指令跟随

用户主要通过自然语言指令，即文本提示[62, 63]，与生成式AI模型进行交互。模型指令调整旨在增强AI模型遵循提示的能力。这种在遵循指令方面的改进能力使模型能够生成更接近于对自然语言查询的人类响应的输出。我们从回顾大型语言模型（LLMs）和诸如DALL·E 3之类的文本到图像模型的指令跟随技术开始讨论。为了增强文本到视频模型遵循文本指令的能力，Sora采用了类似于DALL·E 3的方法。该方法涉及训练描述性字幕器并使用字幕器生成的数据进行微调。由于指令调整，Sora能够适应广泛的用户请求，确保对指令中的细节给予细致的注意，并生成精确满足用户需求的视频。

3.4.1 大型语言模型

LLMs遵循指令的能力已被广泛探索[64, 65, 66]。这种能力使LLMs能够阅读、理解并适当响应描述未见任务的指令，而无需示例。通过在格式化为指令的任务混合物上对LLMs进行微调[64, 66]，获得并增强了遵循提示的能力，这称为指令调整。Wei等人[65]展示了经过指令调整的模型能够更准确地遵循指令。LLMs在未见任务上的表现显著优于未调整的模型。指令遵循能力将LLMs转变为通用任务解决器，标志着AI发展历史上的一个范式转变。

3.4.2 文本到图像

DALL·E 3中的指令遵循通过一种假设为基础的标题改进方法来解决，即模型训练所依赖的文本-图像对的质量决定了最终文本到图像模型的性能。数据的低质量，特别是噪声数据的普遍存在和省略大量视觉信息的简短标题，导致了许多问题，如忽略关键词和词序，以及误解用户意图。标题改进方法通过用详细、描述性的标题重新标注现有图像来解决这些问题。该方法首先训练一个图像标题生成器，这是一个视觉-语言模型，用于生成精确和描述性的图像标题。然后，使用标题生成器生成的描述性图像标题来微调文本到图像模型。具体来说，DALL·E 3遵循对比标题生成器（CoCa）的方法，联合训练一个图像标题生成器，该生成器具有CLIP架构和一个语言模型目标。这个图像标题生成器结合了一个图像编码器、一个单模态文本编码器用于提取语言信息，以及一个多模态文本解码器。它首先使用单模态图像和文本嵌入之间的对比损失，然后是多模态解码器输出的标题生成损失。进一步在详细描述图像的主要对象、周围环境、背景、文本、风格和颜色方面进行微调后，图像标题生成器能够为图像生成详细的描述性标题。文本到图像模型的训练数据集是由图像标题生成器生成的重新标注数据集和真实人类编写的数据混合而成，以确保模型能够捕捉用户输入。这种图像标题改进方法引入了一个潜在问题：实际用户提示与训练数据中的描述性图像描述之间的不匹配。DALL·E 3通过上采样来解决这个问题，其中LLMs被用来将简短的用户提示重写为详细且冗长的指令。这确保了模型在推理时接收到的文本输入与模型训练中的输入一致。

3.4.3 文本到视频

为了增强指令遵循能力，Sora采用了类似的标题改进方法。该方法首先通过训练一个能够为视频生成详细描述的视频标题生成器来实现。然后，将这个视频标题生成器应用于训练数据中的所有视频，以生成高质量的（视频，描述性标题）对，这些对被用来微调Sora，以提高其指令遵循能力。Sora的技术报告没有透露视频标题生成器是如何训练的。鉴于视频标题生成器是一个视频到文本的模型，构建它有许多方法。一种直接的方法是利用CoCa架构进行视频标题生成，通过取视频的多个帧并将每个帧输入到图像编码器，称为VideoCoCa。VideoCoCa基于CoCa，并重用预训练的图像编码器权重，独立地应用于采样的视频帧。结果产生的帧令牌嵌入被展平并连接成一个长序列的视频表示。这些展平的帧令牌随后被一个生成性池化器和一个对比性池化器处理，这两者与对比损失和标题生成损失一起联合训练。构建视频标题生成器的其他替代方法包括mPLUG-2、GIT、FrozenBiLM等。最后，为了确保用户提示与训练数据中的描述性标题格式一致，Sora执行了一个额外的提示扩展步骤，其中GPT-4V被用来将用户输入扩展为详细的描述性提示。

3.4.4 讨论

对于Sora来说，遵循指令的能力对于生成忠实于用户意图的、包含复杂场景的一分钟长视频至关重要。根据Sora的技术报告，这种能力是通过上述方法获得的。通过开发一个能够生成长而详细的标题

的字幕器，然后用这些标题来训练模型。然而，收集用于训练此类字幕器的数据的过程是未知的，且可能需要大量劳动，因为它可能需要视频的详细描述。此外，描述性视频字幕器可能会虚构视频的重要细节。我们认为，如何改进视频字幕器值得进一步研究，并且对于增强文本到图像模型的遵循指令能力至关重要。

3.5 提示工程

提示工程指的是设计和完善提供给AI系统的输入的过程，特别是在生成模型的背景下，以实现特定的或优化的输出。提示工程的艺术和科学涉及以一种引导模型产生尽可能准确、相关和连贯响应的方式来构建这些输入。

3.5.1 文本提示

文本提示工程对于指导文本到视频模型（例如，Sora）生成在视觉上引人注目同时精确满足用户规格的视频至关重要。这涉及到制定详细描述，以指导模型有效地弥合人类创造力和AI执行能力之间的差距。Sora的提示涵盖了广泛的场景。最近的工作（例如，VoP、Make-A-Video和Tune-A-Video）展示了提示工程如何利用模型的自然语言理解能力来解码复杂指令，并将它们渲染成连贯、生动和高质量的视频叙述。如图15所示，“一个穿着时尚的女人走在霓虹灯照亮的东京街道上...”是一个精心制作的文本提示，它确保Sora生成的视频与预期视觉良好对齐。提示工程的质量取决于对词语的仔细选择、提供的细节的具体性，以及理解它们对模型输出的影响。例如，图15中的提示详细指定了场景的动作、设置、角色外观，甚至是期望的情绪和氛围。

图15：一个关于文本到视频生成的提示工程案例研究，使用颜色编码来划分创造过程。蓝色高亮的文本描述了Sora生成的元素，如对一个时尚女性的描绘。相比之下，黄色的文本强调了模型对动作、设置和角色外观的解释，展示了一个精心制作的提示是如何转化为生动和动态的视频叙述的。

3.5.2 图像提示

图像提示作为即将生成视频的内容和其他元素（如角色、设置和情绪）的视觉锚点。此外，文本提示可以指导模型通过添加运动层、互动和叙事进展等来使这些元素动起来，从而使静态图像变得生动。通过利用视觉和文本信息，图像提示允许Sora将静态图像转换为动态、叙事驱动的视频。在图16中，我们展示了使用DALL·E生成的图像提示Sora的AI生成视频的例子，如“一个戴贝雷帽和高领衫的柴犬”、“一个独特的怪物家庭”、“形成‘SORA’字样的云”和“冲浪者在一个历史悠久的大厅内导航巨浪”。这些例子展示了通过向Sora提示DALL·E生成的图像可以实现什么。

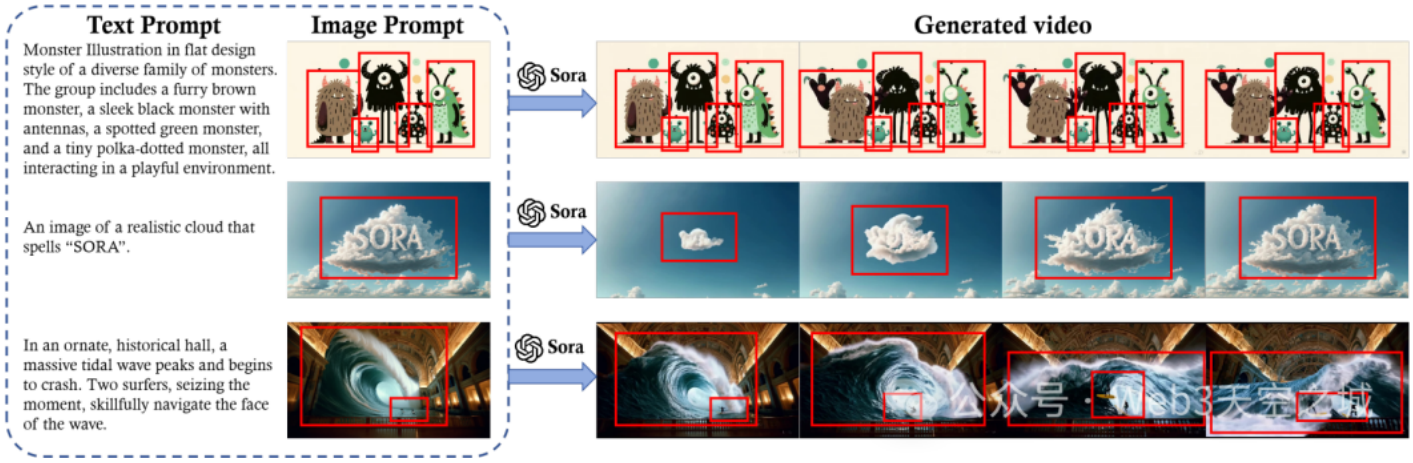


图16：这个例子展示了图像提示指导Sora的文本到视频模型生成的过程。红色框体视觉上锚定了每个场景的关键元素——不同设计的怪物、拼写“SORA”的云形成和在华丽大厅面对巨大海浪的冲浪者。

3.5.3 视频提示

如[82, 83]所示，视频提示也可以用于视频生成。最近的工作（例如，Moonshot和Fast-Vid2Vid）显示，好的视频提示需要具体且灵活。这确保模型在特定目标上获得清晰的指导，如特定对象和视觉主题的描绘，并且还允许在最终输出中进行富有想象力的变化。例如，在视频扩展任务中，提示可以指定方向（向前或向后时间）和扩展的上下文或主题。在图17(a)中，视频提示指导Sora向时间倒退扩展视频，以探索导致原始起点的事件。在通过视频提示进行视频到视频编辑时，如图17(b)所示，模型需要清楚地理解所需的转变，如改变视频的风格、设置或氛围，或改变细微方面，如照明或情绪。在图17(c)中，提示指导Sora连接视频，同时确保不同场景中的对象之间平滑过渡。

3.5.4 讨论

提示工程允许用户指导AI模型生成与他们的意图一致的内容。例如，文本、图像和视频提示的综合使用使Sora能够创建的内容不仅在视觉上引人注目，而且与用户的期望和意图良好对齐。尽管之前的提示工程研究主要集中在LLMs和LVMs的文本和图像提示上，但我们预计对视频生成模型的视频提示将会越来越受到关注。

3.6 可信度

随着ChatGPT、GPT4-V和Sora等复杂模型的快速发展，这些模型的能力得到了显著增强。这些发展为提高工作效率和推动技术进步做出了重大贡献。然而，这些进步也引发了对这些技术潜在滥用的担忧，包括生成假新闻、隐私泄露和伦理困境。因此，在大型模型中的可信度问题已经从学术界和工业界获得了广泛关注，成为当代研究讨论的焦点。

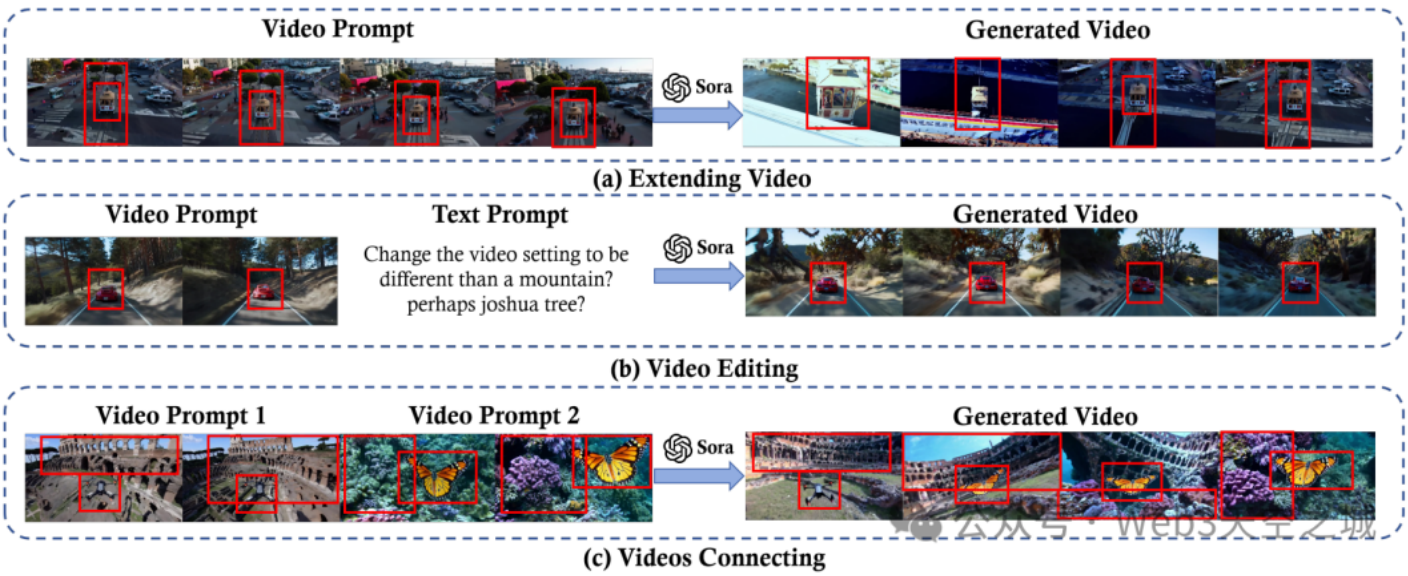


图17：这些示例展示了Sora模型的视频提示技术：(a)视频扩展，模型将序列向原始素材的反方向外推；(b)视频编辑，根据文本提示，特定元素如场景被转换；以及(c)视频连接，两个不同的视频提示被无缝融合以创建连贯的叙述。每个过程都由一个视觉锚点指导，用红色框标记，确保生成的视频内容连续性和精确性。

3.6.1 安全问题

一个主要关注点是模型的安全性，特别是其抵抗滥用和所谓的“越狱”攻击的能力，用户试图利用漏洞生成禁止或有害内容[96, 97, 98, 99, 100, 101, 102, 103, 104, 105]。例如，AutoDAN[103]，一种基于梯度技术的新型可解释对抗攻击方法，被引入以实现系统绕过。在最近的一项研究中，研究人员探讨了大型语言模型(LLMs)难以抵抗越狱攻击的两个原因：目标冲突和泛化不匹配[106]。除了文本攻击外，视觉越狱也威胁到多模态模型的安全（例如，GPT-4V[90]和Sora[3]）。最近的一项研究[107]发现，大型多模态模型更容易受到攻击，因为额外的视觉输入的连续性和高维性使其对对抗攻击更加脆弱，代表了扩大的攻击面。

3.6.2 其他利用

由于大型基础模型（例如，ChatGPT[89]和Sora[3]）的训练数据集规模和训练方法，这些模型的真实性需要得到增强，因为相关问题如幻觉已被广泛讨论[108]。在这个上下文中，幻觉指的是模型倾向于生成看似令人信服但是无根据或虚假的回应[96]。这一现象提出了关于模型输出可靠性和可信度的关键问题，需要采取全面的方法来评估和解决这一问题。已有大量研究致力于从各个角度剖析幻觉问题。这包括旨在评估不同模型和场景中幻觉的程度和性质的努力[109, 96, 110, 111]。这些评估提供了宝贵的见解，说明了幻觉如何以及为什么发生，为制定减少其发生的策略奠定了基础。同时，大量研究集中于设计和实施方法以减少这些大型模型中的幻觉[112, 113, 114]。

另一个关于可信度的重要方面是公平性和偏见。开发不会传播或加剧社会偏见的模型的重要性是一个至关重要的关注点。这一优先事项源于这样的认识：这些模型中编码的偏见可以加强现有的社会不平等，导致歧视性结果。在这一领域的研究，如Gallegos等人[115]、张等人[116]、梁等人[117]和Friedrich等人[118]的工作，致力于这些固有偏见的细致识别和纠正。目标是培养公平运作的模型，公

平对待所有个体，不对种族、性别或其他敏感属性持有偏见。这不仅涉及检测和减轻数据集中的偏见，还包括设计能够主动对抗这些偏见传播的算法[119, 120]。

隐私保护成为部署这些模型时的另一个基础支柱。在数据隐私关切日益升高的时代，保护用户数据的重点从未如此关键。公众对个人数据处理方式的日益关注和担忧促使对大型模型进行了更严格的评估。这些评估关注模型保护用户数据的能力，确保个人信息保持机密，不会无意中泄露。Mireshghallah等人[121]、Plant等人[122]和Li等人[123]的研究代表了推进保护隐私的方法和技术的努力。

3.6.3 对齐

在应对这些挑战时，确保大型模型的可信度已成为研究人员的主要关切之一[124, 96, 99, 125]。最重要的技术之一是模型对齐[125, 126]，它指的是确保模型的行为和输出与人类设计者的意图和伦理标准一致的过程和目标。这涉及技术的发展、其道德责任和社会价值。在大型语言模型(LLMs)领域，使用人类反馈的强化学习(RLHF)[127, 128]方法已被广泛应用于模型对齐。这种方法结合了强化学习(RL)和直接人类反馈，允许模型更好地与人类的期望和标准对齐，以理解和执行任务。

3.6.4 讨论

从Sora（特别是其技术报告）中，我们总结了一些有见地的发现，这些发现可能为未来的工作提供有益的指导：

模型和外部安全的综合保护：随着模型变得更强大，特别是在内容生成方面，确保它们不被滥用来生产有害内容（如仇恨言论[129]和虚假信息[92, 91]）已成为一个严峻的挑战。除了对模型本身进行调整外，外部安全保护同样重要。这包括内容过滤和审查机制、使用权限和访问控制、数据隐私保护，以及在透明度和可解释性方面的增强。例如，OpenAI现在使用检测分类器来判断给定视频是否由Sora生成[130]。此外，还部署了文本分类器来检测潜在有害的文本输入[130]。

多模态模型的安全挑战：多模态模型，如文本到视频模型Sora，由于其理解和生成各种类型内容（文本、图像、视频等）的能力，给安全带来额外的复杂性。多模态模型可以以各种形式生产内容，增加了滥用和版权问题的方式和范围。由于多模态模型生成的内容更加复杂和多样，传统的内容验证和真实性方法可能不再有效。这要求开发新的技术和方法来识别和过滤这些模型生成的有害内容，增加了监管和管理的难度。

跨学科合作的需求：确保模型的安全不仅是一个技术问题，还需要跨学科合作。为了应对这些挑战，来自各个领域的专家，如法律[131]和心理学[132]，需要共同努力制定适当的规范（例如，什么是安全的，什么是不安全的？）、政策和技术解决方案。跨学科合作的需求显著增加了解决这些问题的复杂性。

4 应用

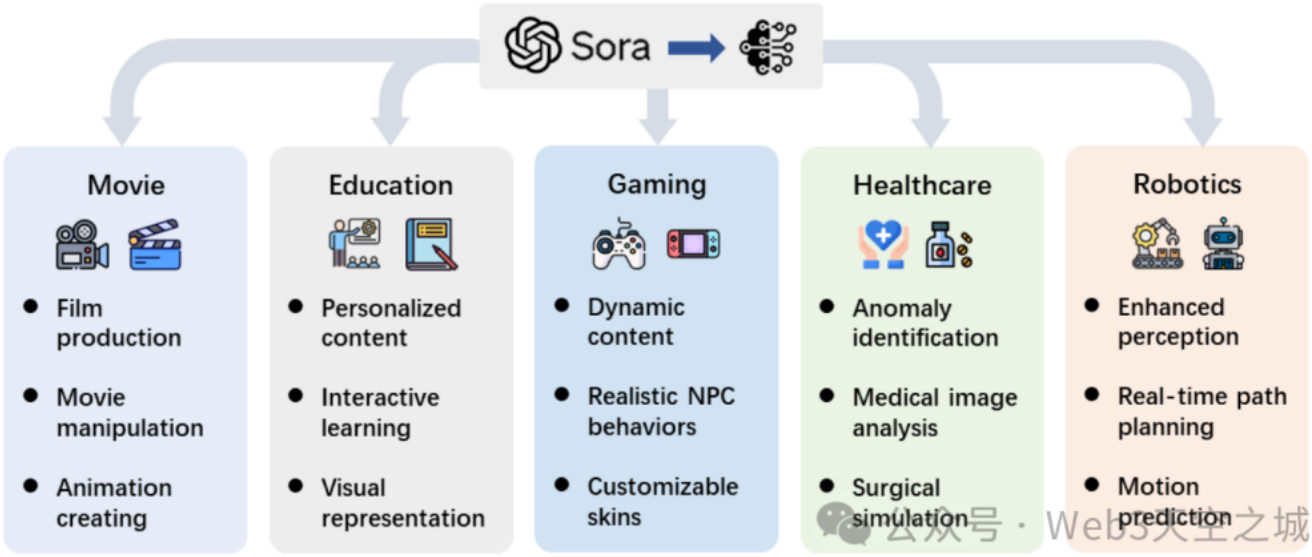


图18：Sora的应用。

随着Sora等视频扩散模型作为前沿技术的兴起，它们在不同研究领域和行业中的采用正在迅速加速。这项技术的影响远远超出了简单的视频创作，为从自动化内容生成到复杂决策过程的任务提供了变革性的潜力。在本节中，我们将深入探讨视频扩散模型当前的应用情况，重点介绍Sora不仅展示了其能力，而且还革新了解决复杂问题方法的关键领域。我们旨在为实际部署场景提供一个广阔的视角（见图18）。

4.1 电影

传统上，创作电影杰作是一个艰巨且昂贵的过程，通常需要数十年的努力、尖端设备和大量的财务投资。然而，先进视频生成技术的出现预示着电影制作的新时代，一个从简单文本输入自主制作电影的梦想正在成为现实的时代。研究人员通过将视频生成模型扩展到电影制作领域，MovieFactory[133]应用扩散模型从ChatGPT[89]生成的详细剧本生成电影风格的视频，代表了一次重大飞跃。在后续中，MobileVidFactory[134]可以仅根据用户提供的简单文本自动生成垂直移动视频。Vlogger[135]使用户能够组成一分钟长的视频博客。这些发展，以Sora轻松生成引人入胜的电影内容的能力为典范，标志着电影制作民主化的关键时刻。它们展示了一个未来的画面，任何人都可以成为电影制作者，显著降低了进入电影行业的门槛，并引入了将传统叙事与AI驱动的创造力相结合的电影制作新维度。这些技术的影响不仅仅是简化，它们承诺将重塑电影制作的格局，使其在不断变化的观众偏好和分发渠道面前变得更加易于访问和多样化。

4.2 教育

教育内容的格局长期以来一直被静态资源所主导，尽管它们具有价值，但往往无法满足当今学生多样化的需求和学习风格。视频扩散模型站在教育革命的前沿，提供了前所未有的机会，以显著增强学习者的参与度和理解力，定制和动态化教育材料。这些先进技术使教育者能够将文本描述或课程大纲转化为针对个别学习者的特定风格和兴趣量身定制的动态、引人入胜的视频内容[136, 137, 138, 139]。此外，图像到视频编辑技术[140, 141, 142]提供了将静态教育资产转换为互动视频的创新途径，从而支持一系列学习偏好，并有可能增加学生的参与度。通过将这些模型整合到教育内容创作中，教育者可以

就广泛的主题制作视频，使复杂概念对学生更加易于理解和吸引人。Sora在教育领域的应用体现了这些技术的变革潜力。这种转向个性化、动态教育内容的转变预示着教育的新时代。

4.3 游戏

游戏行业不断寻求推动现实主义和沉浸感的界限，然而传统游戏开发经常与预渲染环境和脚本事件的限制作斗争。扩散模型生成的动态、高保真视频内容和实时效果的真实声音，承诺克服现有限制，为开发者提供创建响应玩家行动和游戏事件的不断发展的游戏环境的工具[143, 144]。这可能包括生成变化的天气条件、变换的景观，甚至即兴创造全新的设置，使游戏世界更加沉浸和响应。一些方法[145, 146]还从视频输入中合成真实的冲击声音，增强游戏音频体验。通过将Sora整合到游戏领域，可以创造前所未有的沉浸式体验，吸引并吸引玩家。游戏的开发、玩法和体验将得到创新，同时为叙事、互动和沉浸打开新的可能性。

4.4 医疗保健

尽管具有生成能力，视频扩散模型在理解和生成复杂视频序列方面表现出色，使它们特别适合识别身体内的动态异常，如早期细胞凋亡[147]、皮肤病变进展[148]和不规则人体运动[149]，这对于早期疾病检测和干预策略至关重要。此外，像MedSegDiffV2[150]和[151]这样的模型利用变压器的力量，以前所未有的精度对医学图像进行分割，使临床医生能够在各种成像模式中准确地定位感兴趣的区域。将Sora整合到临床实践中，不仅承诺改进诊断过程，还提供基于精确医学成像分析的个性化患者护理。然而，这种技术整合带来了自身的一系列挑战，包括需要强大的数据隐私措施和解决医疗保健中的伦理考虑。

4.5 机器人

视频扩散模型在机器人学中扮演着重要角色，展示了一个新时代，其中机器人可以生成和解释复杂视频序列，以增强感知[152, 153]和决策[154, 155, 156]。这些模型为机器人解锁了新的能力，使它们能够与环境互动并以前所未有的复杂性和精确度执行任务。将网络规模的扩散模型引入机器人学[152]展示了利用大规模模型增强机器人视觉和理解的潜力。潜在扩散模型用于语言指导的视频预测[157]，通过预测视频格式中动作的结果，使机器人能够理解和执行任务。此外，对于机器人研究依赖模拟环境的问题，已通过能够创建高度逼真视频序列的视频扩散模型得到了创新性的解决[158, 159]。这使得能够为机器人生成多样化的训练场景，减轻了现实世界数据稀缺所带来的限制。我们相信，像Sora这样的技术整合到机器人领域，预示着突破性的发展。通过利用Sora的力量，机器人的未来正迎来前所未有的进步，机器人能够无缝地导航和与环境互动。

5 讨论

Sora展现出了精确理解和执行人类复杂指令的非凡才能。这个模型擅长创建具有各种角色的详细视频，所有这些都设置在精心制作的环境中。Sora的一个特别令人印象深刻的属性是其能够制作长达一分钟的视频，同时确保一致且引人入胜的叙事。这标志着相比之前专注于更短视频片段的尝试，Sora的扩展序列展现了清晰的叙事流程，并从头到尾保持视觉一致性。此外，Sora通过生成捕捉复杂动作和互动的更长视频序列，超越了早期模型只能处理短片和基本图像的限制。这一进步在AI驱动的创意

工具中标志着重大步伐，使用户能够将书面故事转化为生动的视频，达到以前无法实现的细节和复杂度水平。

5.1 限制

物理现实性的挑战。作为一个模拟平台，Sora展现出一系列限制，这些限制削弱了其准确描绘复杂场景的有效性。最重要的是它在处理复杂场景中的物理原理时的不一致性，导致无法准确复制特定示例的因果关系。例如，吃掉一部分饼干可能不会产生相应的咬痕，说明系统偶尔偏离物理的可能性。这个问题扩展到运动的模拟，其中Sora生成的运动挑战了现实的物理建模，如物体的不自然变化或椅子等刚性结构的错误模拟，导致不现实的物理互动。当模拟物体和角色之间的复杂互动时，挑战进一步增加，偶尔产生倾向于幽默的结果。

空间和时间的复杂性。Sora偶尔误解与给定提示中的物体和角色的放置或排列相关的指令，导致对方向的混淆（例如，将左右弄反）。此外，它在保持事件的时间准确性方面面临挑战，特别是在遵循指定的摄像机移动或序列时。这可能导致偏离场景预期的时间流。在涉及多个角色或元素的复杂场景中，Sora倾向于插入不相关的动物或人物。这样的添加可以显著改变最初设想的场景构成和氛围，偏离计划的叙事或视觉布局。这个问题不仅影响模型准确重现特定场景或叙事的能力，也影响其生成与用户期望和生成输出的连贯性紧密对齐的内容的可靠性。

人机交互（HCI）的限制。Sora在视频生成领域显示出潜力，但在HCI方面面临重大限制。这些限制主要体现在用户系统互动的连贯性和效率上，特别是在对生成内容进行详细修改或优化时。例如，用户可能发现很难精确指定或调整视频中特定元素的呈现，如动作细节和场景转换。此外，Sora在理解复杂的语言指令或捕捉微妙的语义差异方面的限制可能导致在视频内容方面，若不能完全满足用户的期望或需求，这些不足之处限制了Sora在视频编辑和增强方面的潜力，也影响了用户体验的整体满意度。

使用限制。关于使用限制，OpenAI尚未为Sora公开访问设定具体的发布日期，强调在广泛部署之前对安全性和准备情况采取谨慎的态度。这表明Sora可能仍需要在安全性、隐私保护和内容审查等领域进行进一步的改进和测试。此外，目前Sora只能生成长达一分钟的视频，根据已发布的案例，大多数生成的视频仅有几十秒长。这一限制限制了其在需要较长内容展示的应用中的使用，例如详细的教学视频或深入的故事讲述。这一限制减少了Sora在内容创作中的灵活性。

5.2机遇

学术界。(1) OpenAI推出Sora标志着向鼓励更广泛的AI社区深入探索文本到视频模型的战略转变，利用扩散和变换器技术。这一举措旨在将焦点转向直接从文本描述创建高度复杂和细腻视频内容的潜力，这是一个有望彻底改变内容创作、讲故事和信息分享的前沿。(2) 以其原生大小对Sora进行训练的创新方法，与传统的调整大小或裁剪方法相反，为学术界提供了突破性的灵感。它通过强调使用未修改数据集的好处，开辟了新的路径，这导致了更高级的生成模型的创建。

行业。(1) Sora目前的能力预示着视频模拟技术进步的有希望的道路，突出了在物理和数字领域内显著增强现实感的潜力。Sora通过文本描述创建高度逼真环境的前景为内容创作提供了充满希望的未来。这一潜力扩展到了革命性地改变游戏开发，提供了一个未来的窥视，其中沉浸式生成的世界可以以前所未有的便捷性和准确性被创造。(2) 公司可以利用Sora生产迅速适应市场变化的广告视频，并创建定

制化的营销内容。这不仅降低了生产成本，还增强了广告的吸引力和效果。Sora仅凭文本描述就能生成高度逼真的视频内容的能力，可能会彻底改变品牌与其受众互动的方式，允许创造沉浸式和引人入胜的视频，以前所未有的方式捕捉其产品或服务的精髓。

社会。(1) 虽然利用文本到视频技术替代传统电影制作的前景仍然遥远，但Sora和类似平台对社交媒体上的内容创作具有变革性的潜力。当前视频长度的限制并未减少这些工具在使高质量视频制作普及化方面可以产生的影响，使个人能够在无需昂贵设备的情况下制作引人入胜的内容。这代表了向赋予像TikTok和Reels这样的平台上的内容创作者更多权力的重大转变，开启了创意和参与的新时代。(2) 编剧和创意专业人士可以使用Sora将书面剧本转换为视频，帮助他们更好地展示和分享他们的创意概念，甚至制作短片和动画。从剧本创建详细、生动的视频的能力可以从根本上改变电影制作和动画的前期制作过程，提供了一个窥视未来的故事讲述者如何提案、开发和完善他们叙述的方式。这项技术为脚本开发提供了一种更动态和互动的形式，其中想法可以实时可视化和评估，为创意和协作提供了强大的工具。(3) 记者和新闻机构也可以利用Sora快速生成新闻报道或解释性视频，使新闻内容更加生动和吸引人。这可以显著增加新闻报道的覆盖范围和观众参与度。通过提供一个可以模拟逼真环境和场景的工具，Sora为视觉叙事提供了强大的解决方案。使记者能够通过吸引人的视频传达以前难以制作或成本高昂的复杂故事。总之，Sora在跨越市场营销、新闻业和娱乐业的内容创作方面革命性的潜力是巨大的。

6 结论

我们提供了对Sora的全面审查，以帮助开发者和研究人员研究Sora的能力和相关工作。该审查基于我们对已发布的技术报告的调查和基于现有文献的逆向工程。当Sora的API可用且有关Sora的更多细节被揭露时，我们将继续更新本文。我们希望这篇综述论文能为开源研究社区提供宝贵的资源，并为社区在不久的将来共同开发一个开源版本的Sora奠定基础，以在AIGC时代民主化视频自动创作。为了实现这一目标，我们邀请在所有方面进行讨论、建议和合作。

参考文献

(注：论文英文名可见原文)

[1] OpenAI, “Chatgpt: 获取即时答案，寻找创意灵感，学习新事物。” https://openai.com/chatgpt, 2022。

[2] OpenAI, “Gpt-4技术报告”，2023。

[3] OpenAI, “Sora: 从文本创建视频。” https://openai.com/sora, 2024。

[4] W. Peebles 和 S. Xie, “使用变压器的可扩展扩散模型”，在IEEE/CVF国际计算机视觉会议论文集中，第4195-4205页，2023。

[5] A. A. Efros 和 T. K. Leung, “通过非参数采样的纹理合成”，在第七届IEEE国际计算机视觉会议论文集中，第2卷，第1033-1038页，IEEE，1999。

[6] P. S. Heckbert, “纹理映射概述”，IEEE计算机图形学及应用，第6卷，第11期，第56-67页，1986。

[7] I. J. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, 和 Y. Bengio, “生成对抗网络”，arXiv，2014。

[8] D. P. Kingma 和 M. Welling, “自动编码变分贝叶斯”，arXiv预印本arXiv:1312.6114，2013。

[9] L. Dinh, D. Krueger, 和 Y. Bengio, “Nice: 非线性独立成分估计”，arXiv预印本arXiv:1410.8516，2014。

[10] Y. Song 和 S. Ermon, “通过估计数据分布的梯度进行生成建模”，在神经信息处理系统进展中，第32卷，2019。

[11] Y. Cao, S. Li, Y. Liu, Z. Yan, Y. Dai, P. S. Yu, 和 L. Sun, “AI生成内容的全面调查（AIGC）：从GAN到ChatGPT的生成AI历史”，arXiv预印本arXiv:2303.04226，2023。

[12] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. u. Kaiser, 和 I. Polosukhin, “只需注意力”，在神经信息处理系统进展中（I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, 和 R. Garnett, 编辑），第30卷，Curran Associates, Inc., 2017。

[13] J. Devlin, M.-W. Chang, K. Lee, 和 K. Toutanova, “Bert: 深度双向变压器的语言理解预训练”，arXiv预印本arXiv:1810.04805，2018。

[14] A. Radford, K. Narasimhan, T. Salimans, I. Sutskever, 等， “通过生成预训练提高语言理解”，2018。

[15] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, 等， “一幅图像值得16x16个词：变压器在大规模图像识别中的应用”，arXiv预印本arXiv:2010.11929，2020。

[16] Z. Liu, Y. Lin, Y. Cao, H. Hu, Y. Wei, Z. Zhang, S. Lin, 和 B. Guo, “Swin变压器：使用移动窗口的分层视觉变压器”，在IEEE/CVF国际计算机视觉会议论文集中，第10012-10022页，2021。

[17] O. Ronneberger, P. Fischer, 和 T. Brox, “U-net: 用于生物医学图像分割的卷积网络”，在医学图像计算和计算机辅助干预-MICCAI 2015：第18届国际会议，慕尼黑，德国，2015年10月5-9日，论文集，第三部分18，第234-241页，Springer，2015。

[18] A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark, G. Krueger, 和 I. Sutskever, “从自然语言监督中学习可转移的视觉模型”，2021。

[19] R. Rombach, A. Blattmann, D. Lorenz, P. Esser, 和 B. Ommer, “高分辨率图像合成与潜在扩散模型”，在IEEE/CVF计算机视觉和模式识别会议论文集中，第10684-10695页，2022。

[20] M. AI, “Midjourney: 使用AI艺术生成器从文本到图像。” https://www.midjourneyai.ai/en, 2023。

[21] J. Betker, G. Goh, L. Jing, T. Brooks, J. Wang, L. Li, L. Ouyang, J. Zhuang, J. Lee, Y. Guo, 等，“通过更好的标题改善图像生成”，计算机科学。https://cdn.openai.com/papers/dall-e-3.pdf，第2卷，第3页，2023。

[22] P. AI, “Pika是将您的创意付诸行动的从想法到视频的平台。” https://pika.art/home，2023。

[23] R. AI, “Gen-2: 生成AI的下一步。” https://research.runwayml.com/gen2，2023。

[24] X. Zhai, A. Kolesnikov, N. Houlsby, 和 L. Beyer, “扩展视觉变压器”，在IEEE/CVF计算机视觉和模式识别会议论文集中，第12104-12113页，2022。

[25] M. Dehghani, J. Djolonga, B. Mustafa, P. Padlewski, J. Heek, J. Gilmer, A. P. Steiner, M. Caron, R. Geirhos, I. Alabdulmohsin, 等， “将视觉变压器扩展到220亿参数”，在国际机器学习会议论文集中，第7480-7512页，PMLR，2023。

[26] A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark, 等， “从自然语言监督中学习可转移的视觉模型”，在国际机器学习会议论文集中，第8748-8763页，PMLR，2021。

[27] A. Blattmann, T. Dockhorn, S. Kulal, D. Mendelevitch, M. Kilian, D. Lorenz, Y. Levi, Z. English, V. Voleti, A. Letts, 等， “稳定视频扩散：将潜在视频扩散模型扩展到大型数据集”，arXiv预印本arXiv:2311.15127，2023。[28] U. Singer, A. Polyak, T. Hayes, X. Yin, J. An, S. Zhang, Q. Hu, H. Yang, O. Ashual, O. Gafni, D. Parikh, S. Gupta, 和 Y. Taigman, “无需文本视频数据的文本到视频生成”，2022年。

[29] J. Ho, W. Chan, C. Saharia, J. Whang, R. Gao, A. Gritsenko, D. P. Kingma, B. Poole, M. Norouzi, D. J. Fleet 等， “Imagen video: 使用扩散模型生成高清视频”，arXiv预印本arXiv:2210.02303，2022年。

[30] R. Sutton, “苦涩的教训”。http://www.incompleteideas.net/IncIdeas/BitterLesson.html，2019年3月。访问日期：您的访问日期在此。

[31] S. Xie, “关于sora技术报告的看法”。https://twitter.com/sainingxie/status/1758433676105310543，2024年。

[32] A. Van Den Oord, O. Vinyals 等， “神经离散表示学习”，在神经信息处理系统进展中，第30卷，2017年。

[33] K. He, X. Chen, S. Xie, Y. Li, P. Dollár, 和 R. Girshick, “掩蔽自编码器是可扩展的视觉学习者”，在IEEE/CVF计算机视觉与模式识别会议论文集中，第16000-16009页，2022年。

[34] S. Ge, S. Nah, G. Liu, T. Poon, A. Tao, B. Catanzaro, D. Jacobs, J.-B. Huang, M.-Y. Liu, 和 Y. Balaji， “保留你自己的相关性：视频扩散模型的噪声先验”，在IEEE/CVF国际计算机视觉会议论文集中，第22930-22941页，2023年。

[35] A. Sauer, D. Lorenz, A. Blattmann, 和 R. Rombach, “对抗性扩散蒸馏”，arXiv预印本arXiv:2311.17042，2023年。

[36] A. Blattmann, R. Rombach, H. Ling, T. Dockhorn, S. W. Kim, S. Fidler, 和 K. Kreis, “对齐你的潜变量：使用潜在扩散模型的高分辨率视频合成”，在IEEE/CVF计算机视觉与模式识别会议论文集中，第22563-22575页，2023年。

[37] M. Ryoo, A. Piergiovanni, A. Arnab, M. Dehghani, 和 A. Angelova, “Tokenlearner：视频的自适应时空标记化”，在神经信息处理系统进展中，第34卷，第12786-12797页，2021年。

[38] A. Arnab, M. Dehghani, G. Heigold, C. Sun, M. Lučić, 和 C. Schmid, “Vivit：一个视频视觉变换器”，arXiv预印本arXiv:2103.15691，2021年。

[39] L. Beyer, P. Izmailov, A. Kolesnikov, M. Caron, S. Kornblith, X. Zhai, M. Minderer, M. Tschannen, I. Alabdulmohsin, 和 F. Pavetic, “Flexivit：适用于所有补丁大小的模型”，在IEEE/CVF计算机视觉与模式识别会议论文集中，第14496-14506页，2023年。

[40] M. Dehghani, B. Mustafa, J. Djolonga, J. Heek, M. Minderer, M. Caron, A. Steiner, J. Puigcerver, R. Geirhos, I. M. Alabdulmohsin 等，“Patch n’ pack：Navit，一个适用于任何宽高比和分辨率的视觉变换器”，在神经信息处理系统进展中，第36卷，2024年。

[41] M. M. Krell, M. Kosec, S. P. Perez, 和 A. Fitzgibbon, “高效序列打包无交叉污染：加速大型语言模型而不影响性能”，arXiv预印本arXiv:2107.02027，2021年。

[42] H. Yin, A. Vahdat, J. M. Alvarez, A. Mallya, J. Kautz, 和 P. Molchanov, “A-vit：高效视觉变换器的自适应标记”，在IEEE/CVF计算机视觉与模式识别会议论文集中，第10809-10818页，2022年。

[43] D. Bolya, C.-Y. Fu, X. Dai, P. Zhang, C. Feichtenhofer, 和 J. Hoffman, “标记合并：你的vit但更快”，在第十一届国际学习表示会议中，2022年。

[44] M. Fayyaz, S. A. Koohpayegani, F. R. Jafari, S. Sengupta, H. R. V. Joze, E. Sommerlade, H. Pirsia- vash, 和 J. Gall, “自适应标记采样用于高效视觉变换器”，在欧洲计算机视觉会议中，第396-414页，Springer, 2022年。

[45] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, 和 I. Polosukhin, “注意力就是你所需要的”，在神经信息处理系统进展中，第30卷，2017年。

[46] G. Bertasius, H. Wang, 和 L. Torresani, “时空注意力就是你所需要的用于视频理解吗？”，在ICML中，第2卷，第4页，2021年。

[47] L. Yu, J. Lezama, N. B. Gundavarapu, L. Versari, K. Sohn, D. Minnen, Y. Cheng, A. Gupta, X. Gu, A. G. Hauptmann 等，“语言模型胜过扩散-标记器是视觉生成的关键”，arXiv预印本arXiv:2310.05737，2023年。

[48] N. Shazeer, “快速变换器解码：你所需要的只有一个写头”，2019年。

[49] J. Ainslie, J. Lee-Thorp, M. de Jong, Y. Zemlyanskiy, F. Lebrón, 和 S. Sanghai, “Gqa：从多头检查点训练泛化的多查询变换器模型”，arXiv预印本arXiv:2305.13245，2023年。

[50] A. Gu 和 T. Dao, “Mamba：具有选择性状态空间的线性时间序列建模”，arXiv预印本arXiv:2312.00752，2023年。

[51] J. Sohl-Dickstein, E. A. Weiss, N. Maheswaranathan, 和 S. Ganguli, “使用非平衡热力学的深度无监督学习”，arXiv预印本arXiv:1503.03585，2015年。

[52] J. Ho, A. Jain, 和 P. Abbeel, “去噪扩散概率模型”，在神经信息处理系统进展中，第33卷，第6840-6851页，2020年。

[53] Y. Song, J. Sohl-Dickstein, D. P. Kingma, A. Kumar, S. Ermon, 和 B. Poole, “通过随机微分方程的得分基生成模型”，arXiv预印本arXiv:2011.13456，2020年。

[54] F. Bao, S. Nie, K. Xue, Y. Cao, C. Li, H. Su, 和 J. Zhu, “所有都值得一言：用于扩散模型的vit骨干”，在IEEE/CVF计算机视觉与模式识别会议论文集（CVPR）中，2023年。

[55] S. Gao, P. Zhou, M.-M. Cheng, 和 S. Yan, “掩蔽扩散变换器是一个强大的图像合成器”，arXiv预印本arXiv:2303.14389，2023年。

[56] A. Hatamizadeh, J. Song, G. Liu, J. Kautz, 和 A. Vahdat, “Diffit：用于图像生成的扩散视觉变换器”，arXiv预印本arXiv:2312.02139，2023年。

[57] J. Ho 和 T. Salimans, “无分类器扩散指导”，arXiv预印本arXiv:2207.12598，2022年。

[58] T. Salimans 和 J. Ho, “渐进式蒸馏用于扩散模型的快速采样”，arXiv预印本arXiv:2202.00512，2022年。

[59] J. Ho, C. Saharia, W. Chan, D. J. Fleet, M. Norouzi, 和 T. Salimans, “级联扩散模型用于高保真图像生成”，机器学习研究杂志，第23卷，第1期，第2249-2281页，2022年。[60] R. Rombach, A. Blattmann, D. Lorenz, P. Esser, 和 B. Ommer, “使用潜在扩散模型的高分辨率图像合成”，2021年。

[61] D. Podell, Z. English, K. Lacey, A. Blattmann, T. Dockhorn, J. Müller, J. Penna, 和 R. Rombach, “Sdxl: 改进潜在扩散模型以合成高分辨率图像”，arXiv预印本 arXiv:2307.01952，2023年。

[62] T. B. Brown, B. Mann, N. Ryder, M. Subbiah, J. Kaplan, P. Dhariwal, A. Neelakantan, P. Shyam, G. Sastry, A. Askell 等，“语言模型是少数样本学习者”，arXiv，2020年。

[63] K. Zhou, J. Yang, C. C. Loy, 和 Z. Liu, “条件提示学习用于视觉-语言模型”，在 IEEE/CVF 计算机视觉与模式识别会议论文集中，第16816-16825页，2022年。

[64] V. Sanh, A. Webson, C. Raffel, S. H. Bach, L. Sutawika, Z. Alyafeai, A. Chaffin, A. Stiegler, T. L. Scao, A. Raja 等，“多任务提示训练实现零样本任务泛化”，arXiv预印本 arXiv:2110.08207，2021年。

[65] J. Wei, M. Bosma, V. Y. Zhao, K. Guu, A. W. Yu, B. Lester, N. Du, A. M. Dai, 和 Q. V. Le，“微调语言模型是零样本学习者”，arXiv预印本 arXiv:2109.01652，2021年。

[66] L. Ouyang, J. Wu, X. Jiang, D. Almeida, C. Wainwright, P. Mishkin, C. Zhang, S. Agarwal, K. Slama, A. Ray 等，“训练语言模型根据人类反馈执行指令”，在神经信息处理系统进展中，第35卷，第27730-27744页，2022年。

[67] C. Jia, Y. Yang, Y. Xia, Y.-T. Chen, Z. Parekh, H. Pham, Q. Le, Y.-H. Sung, Z. Li, 和 T. Duerig，“通过噪声文本监督扩大视觉和视觉-语言表示学习”，在国际机器学习会议论文集中，第4904-4916页，PMLR，2021年。

[68] J. Yu, Z. Wang, V. Vasudevan, L. Yeung, M. Seyedhosseini, 和 Y. Wu, “Coca: 对比性标题生成器是图像-文本基础模型”，arXiv预印本 arXiv:2205.01917，2022年。

[69] S. Yan, T. Zhu, Z. Wang, Y. Cao, M. Zhang, S. Ghosh, Y. Wu, 和 J. Yu, “视频-文本建模与对比性标题生成器的零样本迁移”，arXiv预印本 arXiv:2212.04979，2022年。

[70] H. Xu, Q. Ye, M. Yan, Y. Shi, J. Ye, Y. Xu, C. Li, B. Bi, Q. Qian, W. Wang 等，“mplug-2: 一个跨文本、图像和视频的模块化多模态基础模型”，arXiv预印本 arXiv:2302.00402，2023年。

[71] J. Wang, Z. Yang, X. Hu, L. Li, K. Lin, Z. Gan, Z. Liu, C. Liu, 和 L. Wang, “Git: 一个用于视觉和语言的生成图像到文本变换器”，arXiv预印本 arXiv:2205.14100，2022年。

[72] A. Yang, A. Miech, J. Sivic, I. Laptev, 和 C. Schmid, “通过冻结的双向语言模型进行零样本视频问答”，在神经信息处理系统进展中，第35卷，第124-141页，2022年。

[73] Y. Li, “零样本提示设计的实用调查，用于上下文学习”，在自然语言处理的大型语言模型会议最近进展论文集中，RANLP, INCOMA Ltd., Shoumen, 保加利亚，2023年。

[74] B. Chen, Z. Zhang, N. Langrené, 和 S. Zhu, “释放大型语言模型中提示工程的潜力：一项全面回顾”，arXiv预印本 arXiv:2310.14735，2023年。

[75] S. Pitis, M. R. Zhang, A. Wang, 和 J. Ba, “大型语言模型的提升提示集合”，2023年。

[76] Y. Hao, Z. Chi, L. Dong, 和 F. Wei, “优化文本到图像生成的提示”，2023年。

[77] S. Huang, B. Gong, Y. Pan, J. Jiang, Y. Lv, Y. Li, 和 D. Wang, “Vop: 文本-视频合作提示调整用于跨模态检索”，2023年。

[78] J. Z. Wu, Y. Ge, X. Wang, W. Lei, Y. Gu, Y. Shi, W. Hsu, Y. Shan, X. Qie, 和 M. Z. Shou, “Tune-a-video: 一次性调整图像扩散模型用于文本到视频生成”，2023年。

[79] T. Lüddecke 和 A. Ecker, “使用文本和图像提示的图像分割”，在 IEEE/CVF 计算机视觉与模式识别会议论文集中，第7086-7096页，2022年6月。

[80] X. Chen, Y. Wang, L. Zhang, S. Zhuang, X. Ma, J. Yu, Y. Wang, D. Lin, Y. Qiao, 和 Z. Liu, “Seine: 用于生成过渡和预测的短到长视频扩散模型”，2023年。

[81] H. Chen, Y. Zhang, X. Cun, M. Xia, X. Wang, C. Weng, 和 Y. Shan, “Videocrafter2: 克服数据限制以实现高质量视频扩散模型”，2024年。

[82] T.-C. Wang, M.-Y. Liu, J.-Y. Zhu, G. Liu, A. Tao, J. Kautz, 和 B. Catanzaro, “视频到视频合成”，2018年。

[83] T.-C. Wang, M.-Y. Liu, A. Tao, G. Liu, J. Kautz, 和 B. Catanzaro, “少样本视频到视频合成”，2019年。

[84] D. J. Zhang, D. Li, H. Le, M. Z. Shou, C. Xiong, 和 D. Sahoo, “Moonshot: 朝向可控视频生成和编辑的多模态条件”，2024年。

[85] L. Zhuo, G. Wang, S. Li, W. Wu, 和 Z. Liu, “Fast-vid2vid: 用于视频到视频合成的空间-时间压缩”，2022年。

[86] P. Liu, W. Yuan, J. Fu, Z. Jiang, H. Hayashi, 和 G. Neubig, “预训练、提示和预测：自然语言处理中提示方法的系统调查”，2021年。

[87] B. Lester, R. Al-Rfou, 和 N. Constant, “规模的力量对于参数高效的提示调整”，在2021年自然语言处理实证方法会议论文集中，第3045-3059页，2021年。

[88] M. Jia, L. Tang, B.-C. Chen, C. Cardie, S. Belongie, B. Hariharan, 和 S.-N. Lim, “视觉提示调整”，在欧洲计算机视觉会议论文集中，第709-727页，Springer，2022年。

[89] OpenAI, “介绍chatgpt”，2023年。

[90] OpenAI, “Gpt-4v(视觉)系统卡片”，2023年。

[91] Y. Huang 和 L. Sun, “在假新闻中利用chatgpt的力量：生成、检测和解释的深入探索”，2023年。

[92] C. Chen 和 K. Shu, “可以检测到由LLM生成的错误信息吗？”，2023年。

[93] Z. Liu, Y. Huang, X. Yu, L. Zhang, Z. Wu, C. Cao, H. Dai, L. Zhao, Y. Li, P. Shu, F. Zeng, L. Sun, W. Liu, D. Shen, Q. Li, T. Liu, D. Zhu, 和 X. Li, “Deid-gpt: 通过gpt-4进行零样本医学文本去识别”，2023年。[94] J. Yao, X. Yi, X. Wang, Y. Gong, 和 X. Xie, “价值支点：将大型语言模型映射到基本人类价值的多维谱系中”，2023。

[95] Y. Huang, Q. Zhang, P. S. Y, 和 L. Sun, “Trustgpt：一个可信赖和负责任的大型语言模型基准”，2023。

[96] L. Sun, Y. Huang, H. Wang, S. Wu, Q. Zhang, C. Gao, Y. Huang, W. Lyu, Y. Zhang, X. Li, Z. Liu, Y. Liu, Y. Wang, Z. Zhang, B. Kailkhura, C. Xiong, C. Xiao, C. Li, E. Xing, F. Huang, H. Liu, H. Ji, H. Wang, H. Zhang, H. Yao, M. Kellis, M. Zitnik, M. Jiang, M. Bansal, J. Zou, J. Pei, J. Liu, J. Gao, J. Han, J. Zhao, J. Tang, J. Wang, J. Mitchell, K. Shu, K. Xu, K.-W. Chang, L. He, L. Huang, M. Backes, N. Z. Gong, P. S. Yu, P.-Y. Chen, Q. Gu, R. Xu, R. Ying, S. Ji, S. Jana, T. Chen, T. Liu, T. Zhou, W. Wang, X. Li, X. Zhang, X. Wang, X. Xie, X. Chen, X. Wang, Y. Liu, Y. Ye, Y. Cao, Y. Chen, 和 Y. Zhao, “Trustllm：大型语言模型中的可信度”，2024。

[97] M. Mazeika, L. Phan, X. Yin, A. Zou, Z. Wang, N. Mu, E. Sakhaee, N. Li, S. Basart, B. Li, D. Forsyth, 和 D. Hendrycks, “Harmbench：一个用于自动化红队操作和强健拒绝的标准化评估框架”，2024。

[98] Y. Wang, H. Li, X. Han, P. Nakov, 和 T. Baldwin, “不回答：一个用于评估LLMs中安全防护的数据集”，2023。

[99] B. Wang, W. Chen, H. Pei, C. Xie, M. Kang, C. Zhang, C. Xu, Z. Xiong, R. Dutta, R. Schaeffer, 等，“Decodingtrust：GPT模型中信任度的全面评估”，arXiv预印本arXiv:2306.11698，2023。

[100] Z. Zhang, L. Lei, L. Wu, R. Sun, Y. Huang, C. Long, X. Liu, X. Lei, J. Tang, 和 M. Huang, “Safetybench：使用多项选择题评估大型语言模型的安全性”，2023。

[101] X. Shen, Z. Chen, M. Backes, Y. Shen, 和 Y. Zhang, “现在就做任何事：在大型语言模型上表征和评估野外越狱提示”，arXiv预印本arXiv:2308.03825，2023。

[102] X. Liu, N. Xu, M. Chen, 和 C. Xiao, “Autodan：在对齐的大型语言模型上生成隐蔽的越狱提示”，arXiv预印本arXiv:2310.04451，2023。

[103] S. Zhu, R. Zhang, B. An, G. Wu, J. Barrow, Z. Wang, F. Huang, A. Nenkova, 和 T. Sun, “Autodan：对大型语言模型的可解释梯度基对抗攻击”，2023。

[104] A. Zhou, B. Li, 和 H. Wang, “针对越狱攻击保护语言模型的强健提示优化”，arXiv预印本arXiv:2401.17263，2024。

[105] X. Guo, F. Yu, H. Zhang, L. Qin, 和 B. Hu, “Cold-attack：用隐蔽性和可控性越狱LLMs”，2024。

[106] A. Wei, N. Haghtalab, 和 J. Steinhardt, “越狱了：LLM安全训练如何失败？”，arXiv预印本arXiv:2307.02483，2023。

[107] Z. Niu, H. Ren, X. Gao, G. Hua, 和 R. Jin, “针对多模态大型语言模型的越狱攻击”，2024。

[108] H. Liu, W. Xue, Y. Chen, D. Chen, X. Zhao, K. Wang, L. Hou, R. Li, 和 W. Peng, “关于大型视觉-语言模型中的幻觉的调查”，2024。

[109] T. Guan, F. Liu, X. Wu, R. Xian, Z. Li, X. Liu, X. Wang, L. Chen, F. Huang, Y. Yacoob, D. Manocha, 和 T. Zhou, “Hallusionbench：一个用于大型视觉-语言模型中纠缠的语言幻觉和视觉错觉的高级诊断套件”，2023。

[110] Y. Li, Y. Du, K. Zhou, J. Wang, W. X. Zhao, 和 J.-R. Wen, “评估大型视觉-语言模型中的对象幻觉”，2023。

[111] Y. Huang, J. Shi, Y. Li, C. Fan, S. Wu, Q. Zhang, Y. Liu, P. Zhou, Y. Wan, N. Z. Gong, 等，“大型语言模型的Metatool基准：决定是否使用工具以及使用哪种工具”，arXiv预印本arXiv:2310.03128，2023。

[112] F. Liu, K. Lin, L. Li, J. Wang, Y. Yacoob, 和 L. Wang, “通过强健指令调整减轻大型多模态模型中的幻觉”，2023。

[113] L. Wang, J. He, S. Li, N. Liu, 和 E.-P. Lim, “通过用标题重写微调大型视觉-语言模型来减轻细粒度幻觉”，在国际多媒体建模会议上，第32-45页，Springer，2024。

[114] Y. Zhou, C. Cui, J. Yoon, L. Zhang, Z. Deng, C. Finn, M. Bansal, 和 H. Yao, “分析和减轻大型视觉-语言模型中的对象幻觉”，arXiv预印本arXiv:2310.00754，2023。

[115] I. O. Gallegos, R. A. Rossi, J. Barrow, M. M. Tanjim, S. Kim, F. Dernoncourt, T. Yu, R. Zhang, 和 N. K. Ahmed, “大型语言模型中的偏见和公平性：一项调查”，arXiv预印本arXiv:2309.00770，2023。

[116] J. Zhang, K. Bao, Y. Zhang, W. Wang, F. Feng, 和 X. He, “ChatGPT对推荐公平吗？评估大型语言模型推荐中的公平性”，arXiv预印本arXiv:2305.07609，2023。

[117] Y. Liang, L. Cheng, A. Payani, 和 K. Shu, “超越检测：揭示辱骂性语言模型中的公平性漏洞”，2023。

[118] F. Friedrich, P. Schramowski, M. Brack, L. Struppek, D. Hintersdorf, S. Luccioni, 和 K. Kersting, “公平扩散：指导文本到图像生成模型关于公平性”，arXiv预印本arXiv:2302.10893，2023。

[119] R. Liu, C. Jia, J. Wei, G. Xu, L. Wang, 和 S. Vosoughi, “通过加强校准减轻语言模型中的政治偏见”，人工智能会议论文集，第35卷，第14857-14866页，2021年5月。

[120] R. K. Mahabadi, Y. Belinkov, 和 J. Henderson, “通过在语料库中建模偏见来实现端到端的偏见缓解”，2020。

[121] N. Mireshghallah, H. Kim, X. Zhou, Y. Tsvetkov, M. Sap, R. Shokri, 和 Y. Choi, “LLMs能保守秘密吗？通过情境完整性理论测试语言模型的隐私影响”，arXiv预印本arXiv:2310.17884，2023。

[122] R. Plant, V. Giuffrida, 和 D. Gkatzia, “你写什么就是什么：在大型语言模型时代保护隐私”，arXiv预印本arXiv:2204.09391，2022。

[123] H. Li, Y. Chen, J. Luo, Y. Kang, X. Zhang, Q. Hu, C. Chan, 和 Y. Song, “大型语言模型中的隐私：攻击、防御和未来方向”，2023。[124] R. Bommasani, D. A. Hudson, E. Adeli, R. Altman, S. Arora, S. von Arx, M. S. Bernstein, J. Bohg, A. Bosselut, E. Brunskill, E. Brynjolfsson, S. Buch, D. Card, R. Castellon, N. Chatterji, A. Chen, K. Creel, J. Q. Davis, D. Demszky, C. Donahue, M. Doumbouya, E. Durmus, S. Ermon, J. Etchemendy, K. Ethayarajh, L. Fei-Fei, C. Finn, T. Gale, L. Gillespie, K. Goel, N. Goodman, S. Grossman, N. Guha, T. Hashimoto, P. Henderson, J. Hewitt, D. E. Ho, J. Hong, K. Hsu, J. Huang, T. Icard, S. Jain, D. Jurafsky, P. Kalluri, S. Karamcheti, G. Keeling, F. Khani, O. Khattab, P. W. Koh, M. Krass, R. Krishna, R. Kuditipudi, A. Kumar, F. Ladhak, M. Lee, T. Lee, J. Leskovec, I. Levent, X. L. Li, X. Li, T. Ma, A. Malik, C. D. Manning, S. Mirchandani, E. Mitchell, Z. Munyikwa, S. Nair, A. Narayan, D. Narayanan, B. Newman, A. Nie, J. C. Niebles, H. Nilforoshan, J. Nyarko, G. Ogut, L. Orr, I. Papadimitriou, J. S. Park, C. Piech, E. Portelance, C. Potts, A. Raghunathan, R. Reich, H. Ren, F. Rong, Y. Roohani, C. Ruiz, J. Ryan, C. Ré, D. Sadigh, S. Sagawa, K. Santhanam, A. Shih, K. Srinivasan, A. Tamkin, R. Taori, A. W. Thomas, F. Tramèr, R. E. Wang, W. Wang, B. Wu, J. Wu, Y. Wu, S. M. Xie, M. Yasunaga, J. You, M. Zaharia, M. Zhang, T. Zhang, X. Zhang, Y. Zhang, L. Zheng, K. Zhou, 和 P. Liang, “关于基础模型的机遇与风险”，2022年。

[125] T. Shen, R. Jin, Y. Huang, C. Liu, W. Dong, Z. Guo, X. Wu, Y. Liu, 和 D. Xiong, “大型语言模型对齐：一项调查”，arXiv预印本arXiv:2309.15025，2023年。

[126] X. Liu, X. Lei, S. Wang, Y. Huang, Z. Feng, B. Wen, J. Cheng, P. Ke, Y. Xu, W. L. Tam, X. Zhang, L. Sun, H. Wang, J. Zhang, M. Huang, Y. Dong, 和 J. Tang, “Alignbench：大型语言模型中文对齐的基准测试”，2023年。

[127] P. Christiano, J. Leike, T. B. Brown, M. Martic, S. Legg, 和 D. Amodei, “基于人类偏好的深度强化学习”，2023年。

[128] T. Yu, Y. Yao, H. Zhang, T. He, Y. Han, G. Cui, J. Hu, Z. Liu, H.-T. Zheng, M. Sun, 和 T.-S. Chua, “RLHF-V：通过细粒度校正人类反馈实现可信赖的MLLMs行为对齐”，2023年。

[129] M. S. Jahan 和 M. Oussalah, “使用自然语言处理进行仇恨言论自动检测的系统综述。”，《神经计算》，第126232页，2023年。

[130] OpenAI, “Sora安全。”https://openai.com/sora#safety，2024年。

[131] Z. Fei, X. Shen, D. Zhu, F. Zhou, Z. Han, S. Zhang, K. Chen, Z. Shen, 和 J. Ge, “Lawbench：大型语言模型的法律知识基准测试”，arXiv预印本arXiv:2309.16289，2023年。

[132] Y. Li, Y. Huang, Y. Lin, S. Wu, Y. Wan, 和 L. Sun, “我思故我在：使用Awarebench对大型语言模型的意识进行基准测试”，2024年。

[133] J. Zhu, H. Yang, H. He, W. Wang, Z. Tuo, W.-H. Cheng, L. Gao, J. Song, 和 J. Fu, “MovieFactory：使用大型生成模型从文本自动创建电影”，arXiv预印本arXiv:2306.07257，2023年。

[134] J. Zhu, H. Yang, W. Wang, H. He, Z. Tuo, Y. Yu, W.-H. Cheng, L. Gao, J. Song, J. Fu, 等，“MobileVidFactory：从文本为移动设备自动生成基于扩散的社交媒体视频”，在第31届ACM国际多媒体会议论文集中，第9371-9373页，2023年。

[135] S. Zhuang, K. Li, X. Chen, Y. Wang, Z. Liu, Y. Qiao, 和 Y. Wang, “Vlogger：让你的梦想成为视频日志”，arXiv预印本arXiv:2401.09414，2024年。

[136] R. Feng, W. Weng, Y. Wang, Y. Yuan, J. Bao, C. Luo, Z. Chen, 和 B. Guo, “CCEdit：通过扩散模型进行创意和可控的视频编辑”，arXiv预印本arXiv:2309.16496，2023年。

[137] J. Xing, M. Xia, Y. Liu, Y. Zhang, Y. Zhang, Y. He, H. Liu, H. Chen, X. Cun, X. Wang, 等，“Make-your-video：使用文本和结构指导定制视频生成”，arXiv预印本arXiv:2306.00943，2023年。

[138] Y. Guo, C. Yang, A. Rao, Y. Wang, Y. Qiao, D. Lin, 和 B. Dai, “AnimateDiff：无需特定调整即可动画化你的个性化文本到图像扩散模型”，arXiv预印本arXiv:2307.04725，2023年。

[139] Y. He, M. Xia, H. Chen, X. Cun, Y. Gong, J. Xing, Y. Zhang, X. Wang, C. Weng, Y. Shan, 等，“Animate-a-story：通过检索增强的视频生成进行讲故事”，arXiv预印本arXiv:2307.06940，2023年。

[140] H. Ni, C. Shi, K. Li, S. X. Huang, 和 M. R. Min, “使用潜在流扩散模型的条件图像到视频生成”，在IEEE/CVF计算机视觉与模式识别会议论文集中，第18444-18455页，2023年。

[141] L. Hu, X. Gao, P. Zhang, K. Sun, B. Zhang, 和 L. Bo, “Animate anyone：一致且可控的图像到视频合成用于角色动画”，arXiv预印本arXiv:2311.17117，2023年。

[142] Y. Hu, C. Luo, 和 Z. Chen, “让它动起来：可控的图像到视频生成与文本描述”，在IEEE/CVF计算机视觉与模式识别会议论文集中，第18219-18228页，2022年。

[143] K. Mei 和 V. Patel, “Vidm：视频隐式扩散模型”，在人工智能AAAI会议论文集中，第37卷，第9117-9125页，2023年。

[144] S. Yu, K. Sohn, S. Kim, 和 J. Shin, “投影潜在空间中的视频概率扩散模型”，在IEEE/CVF计算机视觉与模式识别会议论文集中，第18456-18466页，2023年。

[145] K. Su, K. Qian, E. Shlizerman, A. Torralba, 和 C. Gan, “物理驱动的扩散模型用于从视频合成撞击声音”，在IEEE/CVF计算机视觉与模式识别会议论文集中，第9749-9759页，2023年。

[146] S. Li, W. Dong, Y. Zhang, F. Tang, C. Ma, O. Deussen, T.-Y. Lee, 和 C. Xu, “使用基于编码器的文本反转扩散模型的舞蹈到音乐生成”，arXiv预印本arXiv:2401.17800，2024年。

[147] A. Awasthi, J. Nizam, S. Zare, S. Ahmad, M. J. Montalvo, N. Varadarajan, B. Roysam, 和 H. V. Nguyen, “用于凋亡预测的视频扩散模型”，bioRxiv，第2023-11页，2023年。

[148] A. Bozorgpour, Y. Sadegheih, A. Kazerouni, R. Azad, 和 D. Merhof, “Dermosegdiff：用于皮肤病变勾画的边界感知分割扩散模型”，在PRe预测智能医学国际研讨会论文集中，第146-158页，

Springer，2023年。

[149] A. Flaborea, L. Collorone, G. M. D. di Melendugno, S. D’Arrigo, B. Prenkaj, 和 F. Galasso, “多模态运动条件扩散模型用于基于骨架的视频异常检测”，在IEEE/CVF国际计算机视觉会议论文集中，第10318-10329页，2023年。

[150] J. Wu, R. Fu, H. Fang, Y. Zhang, 和 Y. Xu, “Medsegdiff-v2：基于扩散的医学图像分割与变压器”，arXiv预印本arXiv:2301.11798，2023年。[151] G. J. Chowdary和Z. Yin，“用于医学图像分割的扩散变换器U-Net”，在国际医学图像计算和计算机辅助干预会议上，第622-631页，Springer，2023年。

[152] I. Kapelyukh, V. Vosylius和E. Johns，“Dall-e-bot：将网络规模扩散模型引入机器人技术”，IEEE机器人与自动化信函，2023年。

[153] W. Liu, T. Hermans, S. Chernova和C. Paxton，“Structdiffusion：面向对象的扩散用于新颖对象的语义重排”，在CoRL 2022的语言与机器人学研讨会上，2022年。

[154] M. Janner, Y. Du, J. B. Tenenbaum和S. Levine，“计划与扩散：灵活行为合成的规划”，arXiv预印本arXiv:2205.09991，2022年。

[155] A. Ajay, Y. Du, A. Gupta, J. Tenenbaum, T. Jaakkola和P. Agrawal，“条件生成建模是决策所需的全部吗？”，arXiv预印本arXiv:2211.15657，2022年。

[156] J. Carvalho, A. T. Le, M. Baierl, D. Koert和J. Peters，“运动规划扩散：用扩散模型学习和规划机器人运动”，在2023 IEEE/RSJ国际智能机器人与系统会议（IROS）上，第1916-1923页，IEEE，2023年。

[157] X. Gu, C. Wen, J. Song和Y. Gao，“Seer：用潜在扩散模型进行语言指导的视频预测”，arXiv预印本arXiv:2303.14897，2023年。

[158] Z. Chen, S. Kiami, A. Gupta和V. Kumar，“Genaug：通过生成增强将行为重定向到未见情境”，arXiv预印本arXiv:2302.06671，2023年。

[159] Z. Mandi, H. Bharadhwaj, V. Moens, S. Song, A. Rajeswaran和V. Kumar，“Cacti：一个可扩展的多任务多场景视觉模仿学习框架”，arXiv预印本arXiv:2212.05711，2022年。

[160] T. Chen, L. Li, S. Saxena, G. Hinton和D. J. Fleet，“全景分割图像和视频的通用框架”，在IEEE/CVF国际计算机视觉会议论文集中，第909-919页，2023年。

[161] W. Harvey, S. Naderiparizi, V. Masrani, C. Weilbach和F. Wood，“长视频的灵活扩散建模”，在神经信息处理系统进展中，第35卷，第27953-27965页，2022年。

[162] A. Gupta, S. Tian, Y. Zhang, J. Wu, R. Martín-Martín和L. Fei-Fei，“Maskvit：视频预测的遮罩视觉预训练”，arXiv预印本arXiv:2206.11894，2022年。

[163] W. Hong, M. Ding, W. Zheng, X. Liu和J. Tang，“Cogvideo：通过变换器进行大规模文本到视频生成的预训练”，arXiv预印本arXiv:2205.15868，2022年。

[164] U. Singer, A. Polyak, T. Hayes, X. Yin, J. An, S. Zhang, Q. Hu, H. Yang, O. Ashual, O. Gafni等，“Make-a-video：无文本视频数据的文本到视频生成”，arXiv预印本arXiv:2209.14792，2022

年。

[165] D. Zhou, W. Wang, H. Yan, W. Lv, Y. Zhu和J. Feng，“Magicvideo：用潜在扩散模型高效视频生成”，arXiv预印本arXiv:2211.11018，2022年。

[166] S. Ge, T. Hayes, H. Yang, X. Yin, G. Pang, D. Jacobs, J.-B. Huang和D. Parikh，“用时间不变的VQGAN和时间敏感的变换器生成长视频”，在欧洲计算机视觉会议上，第102-118页，Springer，2022年。

[167] R. Villegas, M. Babaeizadeh, P.-J. Kindermans, H. Moraldo, H. Zhang, M. T. Saffar, S. Castro, J. Kunze和D. Erhan，“Phenaki：从开放域文本描述生成可变长度视频”，arXiv预印本arXiv:2210.02399，2022年。

[168] P. Esser, J. Chiu, P. Atighehchian, J. Granskog和A. Germanidis，“用扩散模型引导的结构和内容视频合成”，在IEEE/CVF国际计算机视觉会议论文集中，第7346-7356页，2023年。

[169] L. Khachatryan, A. Movsisyan, V. Tadevosyan, R. Henschel, Z. Wang, S. Navasardyan和H. Shi，“Text2video-zero：文本到图像扩散模型是零次视频生成器”，arXiv预印本arXiv:2303.13439，2023年。

[170] Z. Luo, D. Chen, Y. Zhang, Y. Huang, L. Wang, Y. Shen, D. Zhao, J. Zhou和T. Tan，“Videofusion：用于高质量视频生成的分解扩散模型”，在IEEE/CVF计算机视觉与模式识别会议论文集中，第10209-10218页，2023年。

[171] A. Jabri, D. Fleet和T. Chen，“可扩展的自适应计算用于迭代生成”，arXiv预印本arXiv:2212.11972，2022年。

[172] L. Lian, B. Shi, A. Yala, T. Darrell和B. Li，“以LLM为基础的视频扩散模型”，arXiv预印本arXiv:2309.17444，2023年。

[173] E. Molad, E. Horwitz, D. Valevski, A. R. Acha, Y. Matias, Y. Pritch, Y. Leviathan和Y. Hoshen，“Dreamix：视频扩散模型是通用视频编辑器”，arXiv预印本arXiv:2302.01329，2023年。

[174] J. H. Liew, H. Yan, J. Zhang, Z. Xu和J. Feng，“Magicedit：高保真和时间连贯的视频编辑”，arXiv预印本arXiv:2308.14749，2023年。

[175] W. Chen, J. Wu, P. Xie, H. Wu, J. Li, X. Xia, X. Xiao和L. Lin，“Control-a-video：用扩散模型进行可控文本到视频生成”，arXiv预印本arXiv:2305.13840，2023年。

[176] W. Chai, X. Guo, G. Wang和Y. Lu，“Stablevideo：文本驱动的一致性感知扩散视频编辑”，在IEEE/CVF国际计算机视觉会议论文集中，第23040-23050页，2023年。

[177] S. Yang, Y. Zhou, Z. Liu和C. C. Loy，“重新渲染视频：零次文本引导的视频到视频翻译”，arXiv预印本arXiv:2306.07954，2023年。

[178] D. Ceylan, C.-H. P. Huang和N. J. Mitra，“Pix2video：使用图像扩散进行视频编辑”，在IEEE/CVF国际计算机视觉会议论文集中，第23206-23217页，2023年。

[179] B. Qin, J. Li, S. Tang, T.-S. Chua和Y. Zhuang，“Instructvid2vid：用自然语言指令进行可控视频编辑”，arXiv预印本arXiv:2305.12328，2023年。

[180] D. Liu, Q. Li, A.-D. Dinh, T. Jiang, M. Shah和C. Xu，“扩散动作分割”，在IEEE/CVF国际计算机视觉会议论文集中，第10139-10149页，2023年。

[181] R. Feng, Y. Gao, T. H. E. Tse, X. Ma和H. J. Chang，“Diffpose：基于视频的人体姿态估计的时空扩散模型”，在IEEE/CVF国际计算机视觉会议论文集中，第14861-14872页，2023年。[182] 余L., 程Y., 宋K., 莱扎马J., 张H., 常H., 郝A.G., 杨M.-H., 郝Y., 艾萨I.等，“Magvit: 掩蔽式生成视频变换器”，在IEEE/CVF计算机视觉与模式识别会议论文集中，第10459-10469页，2023年。

[183] 李Z., 塔克R., 斯纳夫利N., 和霍林斯基A.，“生成图像动态”，arXiv预印本arXiv:2309.07906，2023年。

[184] EasyWithAI，“Zeroscope - AI文本到视频模型。”https://easywithai.com/ai-video-generators/zeroscope/，2023年。

[185] 吉尔达R., 辛格M., 布朗A., 杜瓦尔Q., 阿扎迪S., 拉姆巴特拉S.S., 沙阿A., 尹X., 帕里克D.和米斯拉I.，“Emu视频：通过显式图像条件分解文本到视频生成”，arXiv预印本arXiv:2311.10709，2023年。

[186] 曾Y., 魏G., 郑J., 邹J., 魏Y., 张Y.和李H.，“让像素跳舞：高动态视频生成”，arXiv预印本arXiv:2311.10982，2023年。

[187] 古普塔A., 余L., 宋K., 顾X., 汉M., 费-费L., 艾萨I., 江L.和莱扎马J.，“使用扩散模型的逼真视频生成”，arXiv预印本arXiv:2312.06662，2023年。

[188] 吴B., 莊C.-Y., 王X., 贾Y., 克里希纳库马尔K., 肖T., 梁F., 余L.和瓦贾P.，“Fairy: 快速并行化指导视频到视频合成”，arXiv预印本arXiv:2312.13834，2023年。

[189] 康德拉秋克D., 余L., 顾X., 莱扎马J., 黄J., 霍恩宁R., 亚当H., 阿克巴里H., 阿隆Y., 比罗德卡V.等，“VideoPoet: 一个用于零样本视频生成的大型语言模型”，arXiv预印本arXiv:2312.14125，2023年。

[190] 吴J., 李X., 司C., 周S., 杨J., 张J., 李Y., 陈K., 童Y., 刘Z.等，“通过多模态大型语言模型实现面向语言的视频修复”，arXiv预印本arXiv:2401.10226，2024年。

[191] 巴-塔尔O., 舍弗H., 托夫O., 赫尔曼C., 帕斯R., 扎达S., 埃弗拉特A., 胡尔J., 李Y., 迈克尔T.等，“Lumiere: 一个用于视频生成的时空扩散模型”，arXiv预印本arXiv:2401.12945，2024年。

相关工作

我们在表1中展示了一些与视频生成任务相关的工作。

表1：视频生成概述。

模型名称　年份　基础架构　任务　组织

Imagen Video[29] 2022 扩散 生成 Google

Pix2Seq-D[160] 2022 扩散 分割 Google Deepmind

FDM[161] 2022 扩散 预测 UBC

MaskViT[162] 2022 掩蔽视觉模型 预测 Stanford, Salesforce

CogVideo[163] 2022 自回归 生成 THU

Make-a-video[164] 2022 扩散 生成 Meta

MagicVideo[165] 2022 扩散 生成 ByteDance

TATS[166] 2022 自回归 生成 University of Maryland, Meta

Phenaki[167] 2022 掩蔽视觉模型 生成 Google Brain

Gen-1[168] 2023 扩散 生成, 编辑 RunwayML

LFDM[140] 2023 扩散 生成 PSU, UCSD

Text2video-Zero[169] 2023 扩散 生成 Picsart

Video Fusion[170] 2023 扩散 生成 USAC, Alibaba

PYoCo[34] 2023 扩散 生成 Nvidia

Video LDM[36] 2023 扩散 生成 University of Maryland, Nvidia

RIN[171] 2023 扩散 生成 Google Brain

LVD[172] 2023 扩散 生成 UCB

Dreamix[173] 2023 扩散 编辑 Google

MagicEdit[174] 2023 扩散 编辑 ByteDance

Control-A-Video[175] 2023 扩散 编辑 Sun Yat-Sen University

StableVideo[176] 2023 扩散 编辑 ZJU, MSRA

Tune-A-Video[78] 2023 扩散 编辑 NUS

Rerender-A-Video[177] 2023 扩散 编辑 NTU

Pix2Video[178] 2023 扩散 编辑 Adobe, UCL

InstructVid2Vid[179] 2023 扩散 编辑 ZJU

DiffAct[180] 2023 扩散 动作检测 University of Sydney

DiffPose[181] 2023 扩散 姿态估计 Jilin University

MAGVIT[182] 2023 掩蔽视觉模型 生成 Google

AnimateDiff[138] 2023 扩散 生成 CUHK

MAGVIT V2[47] 2023 掩蔽视觉模型 生成 Google

Generative Dynamics[183] 2023 扩散 生成 Google

VideoCrafter[81] 2023 扩散 生成 Tencent

Zeroscope[184] 2023 - 生成 EasyWithAI

ModelScope 2023 - 生成 Damo

Gen-2[23] 2023 - 生成 RunwayML

Pika[22] 2023 - 生成 Pika Labs

Emu Video[185] 2023 扩散 生成 Meta

PixelDance[186] 2023 扩散 生成 ByteDance

Stable Video Diffusion[27] 2023 扩散 生成 Stability AI

W.A.L.T[187] 2023 扩散 生成 Stanford, Google

Fairy[188] 2023 扩散 生成, 编辑 Meta

VideoPoet[189] 2023 自回归 生成, 编辑 Google

LGVI[190] 2024 扩散 编辑 PKU, NTU

Lumiere[191] 2024 扩散 生成 Google

Sora[3] 2024 扩散 生成, 编辑 OpenAI